

STIL 2023

14th Symposium in Information and Human
Language Technology

Proceedings of the Conference, Vol. 1

December 29, 2023

About the workshop

The Proceedings of the XIV Brazilian Symposium on Information Technology and Human Language (STIL 2023) present the selected papers presented at the event held from September 25 to 29, 2023, in the city of Belo Horizonte, MG, in conjunction with the XII Brazilian Conference on Intelligent Systems (BRACIS 2023), XX National Meeting on Artificial and Computational Intelligence (ENIAC 2023), and VI Brazilian Competition on Knowledge Discovery In Databases (KDD-BR 2023). In this edition, the proceedings compile the selected papers for the Main Conference and the works accepted in its satellite events:

- XIV Brazilian Symposium on Information Technology and Human Language (STIL 2023) received a total of 73 paper submissions, of which 25 were accepted, resulting in an acceptance rate of 34.2%. These works were selected through a double-blind peer review process;
- VIII Jornada de Descrição do Português (JDP 2023) received a total of 21 paper submissions, of which 12 were accepted, resulting in an acceptance rate of 57.1%. These works were selected through a double-blind peer review process;
- VIII Workshop on Scientific Initiation in Information Technology and Human Language (TILic 2023) received a total of 15 paper submissions, of which 13 were accepted, resulting in an acceptance rate of 86.6%. These works were selected through a double-blind peer review process.

Acknowledgments

The Program Committee chairs of STIL and collocated events acknowledge the support to the conference provided by the Brazilian Computer Society (SBC), and the North American Chapter of Association for Computational Linguistics (NAACL). We thank the Program Committees of the XIV Brazilian Symposium in Information and Human Language Technology and Collocated Events for the reviews that they produced. Last but not least, we are grateful to the local organization led by Gisele Pappa (UFMG) and Wagner Meira Jr (UFMG), general chairs of BRACIS 2023.

December 2023

Helena de Medeiros Caseli (UFSCar, Brazil)
Maria José Bocorny Finatto (UFRGS, Brazil)

Program chairs

- **STIL**
Helena Caseli (UFSCar, Brazil)
Maria José Bocorny Finatto (UFRGS, Brazil)
- **JDP**
Cláudia Freitas (ICMC/USP)
Gabriela Wick Pedro (IBICT/NILC)
- **TILIC**
Roney Lira de Sales Santos (UFBA)
Raimundo Santos Moura (UFPI)
Rafael T. Anchiêta (IFPI)

Program Comittee

Adriana Pagano (UFMG)	Leonardo Zilio (UFRGS)
Adriana Silvina Pagano (UFMG)	Lucene Lopes (USP)
Alexandre Rademaker (IBM)	Magali Duran (USP/ICMC)
Aline Evers (UFRGS)	Magnun Madruga (UFMG)
Amanda Rassi (Red. Nota 1000)	Marcela Couto (UFMG)
Ariani Di Felippo (UFSCar)	Marcio Inácio (Univ. de Coimbra)
Arnaldo Candido Junior (UTFPR)	Marcos Lopes (USP)
Celso Ferrarezi (UNIFAL-MG)	Maria Cantoni (UFMG)
Clarissa Xavier (SiDi)	Maria G. Volpe Nunes (USP)
Cláudia Freitas (PUC-Rio)	Maria José Finatto (UFRGS)
Daniela Claro (UFBA)	Marlo Souza (UFBA)
Diana Santos (Univ. of Oslo)	Nathan Hartmann (Itaú Unibanco)
Eloize Rossi Marques Seno (IFSP)	Norton Trevisan Roman (USP)
Eric Laporte (Uni. Gustave Eiffel)	Oto Vale (UFSCar)
Evandro Fonseca (Take Blip)	Paula C. Figueira Cardoso (UFLA)
Guilherme Fromm (UFU)	Raquel Freitag (UFS)
Heliana Mello (UFMG)	Roana Rodrigues (UFS)
Jackson Souza (UFBA)	Rogério Sousa (IFPI)
Joakim Nivre (Uppsala University)	Stella Tagnin (USP)
Jorge Baptista (Univ. do Algarve)	Thiago Pardo (USP/ICMC)
Juliano Desiderato Antonio (UEM)	Tiago Timponi Torrent (UFJF)
Larissa Freitas (UFPEL)	Valeria de Paiva (Topos Institute)

Proceedings organization

Jackson Wilke da Cruz Souza (UFBA, Brazil)
Helena de Medeiros Caseli (UFSCar, Brazil)
Maria José Bocorny Finatto (UFRGS, Brazil)

Contents

STIL

Less is More? Investigating Meta-Learning's Suitability in Sentence Compression for Low-Resource Data

Luis Gustavo C. Rêgo., José Antônio F. de Macêdo, Ticiane L. Coelho da Silva..... 10-19

Predição de transtorno depressivo em redes sociais: BERT supervisionado ou ChatGPT zero-shot?

Wesley Ramos dos Santos, Ivandré Paraboni..... 20-30

Semantic Textual Similarity: In Defense of Wordnet-Based Methods

Eduardo Corrêa Gonçalves..... 31-40

Contextual stance classification using prompt engineering

Felipe Penhorate Carvalho de Fonseca, Ivandré Paraboni, Luciano Antonio Digiampietri..... 41-51

Classificação de gêneros a partir de letras de músicas em português

Matheus Bastos de Oliveira, João Baptista de Oliveira e Souza Filho..... 52-61

Sexismo no Brasil: análise de um Word Embedding por meio de testes baseados em associação implícita

Fernanda Tiemi de S. Taso, Valéria Quadros dos Reis, Fábio Viduani Martinez..... 62-71

Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo "Universal Dependencies"

Emanuel Huber Silva, Thiago Alexandre Salgueiro Pardo, Norton Trevisan Roman..... 72-82

Automated question answering via natural language sentence similarity: Achievements for Brazilian e-commerce platforms

Víctor Jesús Sotelo Chico, Luiz Zucchi, Daniel Ferragut, Rodrigo Caus, Victor Hochgreb de Freitas, Julio Cesar dos Reis..... 83-92

Classificação de Polaridade Orientada aos Alvos de Opinião em Comentários sobre Debate Político em Português

Eloize R. Marques Seno, Fábio S. Igarashi Anno, Lucas Lazarini, Helena M. Caseli..... 93-102

How Good Is ChatGPT For Detecting Hate Speech In Portuguese?

Amanda S. Oliveira, Thiago C. Cecote, Pedro H. L. Silva, Jadson C. Gertrudes, Vander L. S. Freitas, Eduardo J. S. Luz..... 103-112

When Tweets Get Viral - A Deep Learning Approach for Stance Analysis of Covid-19 Vaccines Tweets by Brazilian Political Elites

Lorena Guadalupe Barberia, Pedro Henrique de Santana Schmalz, Norton Trevisan Roman..... 113-123

The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion <i>Magali S. Duran, Lucelene Lopes, Maria das Graças V. Nunes, Thiago A. S. Pardo</i>	124-133
Explorando variações no tagset e na anotação Universal Dependencies (UD) para Português: Possibilidades e resultados com base no treebank PetroGold <i>Elvis de Souza, Cláudia Freitas</i>	134-143
Yauti: A Tool for Morphosyntactic Analysis of Nheengatu within the Universal Dependencies Framework <i>Leonel Figueiredo de Alencar</i>	144-154
Albertina in Action: An Investigation of its Abilities in Aspect Extraction, Hate Speech Detection, Irony Detection, and Question-Answering <i>Júlia da Rocha Junqueira, Claudio Luis Junior, Félix Leonel V. Silva, Ulisses Brisolara Côrrea, Larissa A. de Freitas</i>	155-164
Studying the Dependence of Embedding Representations on the Target of NLP Tasks <i>Bárbara Stéphanie Neves Oliveira, Ticiane L. Coelho da Silva, José A. F. de Macêdo</i>	165-175
Viés de gênero na tradução automática do GPT-3.5 turbo: avaliando o par linguístico inglês-português <i>Tayane Arantes Soares, Yohan Bonescki Gumiel, Rafael Junqueira, Tácio Gomes, Adriana Pagano</i>	176-185
CDJUR-BR - Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas <i>Maurício Brito, Vlândia Pinheiro, Vasco Furtado, João Araújo Monteiro Neto, Francisco das Chagas Jucá Bomfim, André Câmara Ferreira da Costa, Raquel Silveira</i>	186-195
A call for a research agenda on fair NLP for Portuguese <i>Luiz Fernando F. P. de Lima, Renata Mendes de Araujo</i>	196-201
Avaliação do senso comum em modelos de linguagem através de benchmarks: Desafio de Winograd aplicado ao ChatGPT em português brasileiro <i>Thiago Gomes do Nascimento, Diogo Cortiz</i>	202-207
A Sentiment Analysis Benchmark for Automated Machine Learning Applications and a Proof of Concept in Hate Speech Detection <i>Marília Costa Rosendo Silva, Vitor Augusto de Oliveira, Thiago Alexandre Salgueiro Pardo</i>	208-215
Proposta e Avaliação Linguística de Técnicas de Aumento de Dados <i>Arthur Scalercio, Cláudia Freitas</i>	216-232
Semantic Textual Similarity for Abridging Clinical Notes in Brazilian Electronic Health Records <i>Lucas T. Bandeira, Bernardo S. Consoli, Renata Vieira, Rafael H. Bordin</i>	233-237

Towards analysis on textual inference at ASSIN-2 dataset

Felipe O. da Silva, Giovana M. Craveiro, Vinícius F. da Silva, Vinícius João B. Vanzin.....238-243

Previsão de Utilidade de Avaliações de Produtos Online na Língua Portuguesa Brasileira

Larissa F. S. Britto, Luciano D. S. Pacífico, Teresa B. Ludermit.....244-248

JORNADA DE DESCRIÇÃO DO PORTUGUÊS

Tipologia de fenômenos ortográficos e lexicais em CGU: o caso dos tweets do mercado financeiro

Clarissa Lenina Scandarolli, Ariani Di Felippo, Norton Trevisan Roman, Thiago A. S. Pardo.....249-257

Sinalizadores retórico-discursivos: revisitando a anotação RST no cópua CSTNews

Roana Rodrigues, Jackson Wilke da Cruz Souza, Paula Christina Figueira Cardoso.....258-266

Complexidade textual em narrativas orais produzidas por informantes de diferentes níveis de escolaridade

Juliano Desiderato Antonio.....267-276

Aposições anafóricas e catafóricas no português e sua anotação no esquema Universal Dependencies

Magali Sanches Duran, Maria das Graças Volpe Nunes.....277-286

Um pronome com muitas funções: Descrição e resultados da anotação do pronome -se em um treebank segundo o esquema Universal Dependencies (UD) para Português

Elvis de Souza, Cláudia Freitas.....287-296

A funcionalidade dos adjetivos em dois gêneros discursivos: uma investigação com base nas dependências universais

Andre V. Lopes Coneglian, Adriana Pagano, Carlos Perini.....297-306

Indução Gramatical para o Português: a Contribuição da Informação Mútua para Descoberta de Relações de Dependência

Diego Pedro Gonçalves da Silva, Thiago Alexandre Salgueiro Pardo.....307-316

Características gerais das leis federais brasileiras: um estudo exploratório sob o enfoque da Análise Multidimensional

Carolina Godoi de Faria Marques, Lúcia de Almeida Ferrari.....317-326

Processamento da Linguagem Natural e Complexidade Sintática

Bruna R. da Silva.....327-336

Formal features for a syntactic-semantic classification of predicative adjectives in Brazilian Portuguese

Ryan Marçal Saldanha Magaña Martinez, Oto Araújo Vale.....337-346

Building a Frame-Semantic Model of the Healthcare Domain: Towards the identification of gender-based violence in public health data

Lívia Dutra, Arthur Lorenzi, Lorena Larré, Frederico Belcavello, Ely Matos, Amanda Pestana, Kenneth Brown, Mariana Gonçalves, Victor Herbst, Sofia Reinach, Renato Teixeira, Pedro de Paula, Alessandra Pellini, Cibele Sequeira, Ester Sabino, Fábio Leal, Mônica Conde, Regina Grespan, Tiago Torrent.....347-355

WORKSHOP DE IC EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA

Gramáticas Locais para Reconhecimento de Construções com Verbo Suporte em Português

Luis Enrique Santos Prado Vereau, Juliana Pinheiro Campos Pirovani.....356-360

Anotação do Dataset Multimodal da ReINVenTA

Ana Carolina Loçasso Luz, Gabrielly Braz, Lívia Pádua Ruiz, Mariane de Carvalho Pinto, Frederico Belcavello, Natália Sathler Sigiliano, Tiago Torrent.....361-365

Pipeline para identificação de erros lexicais e geração de sugestões de correção

Luana Q. Garcia, Miguel H. Chinellato, Helena de M. Caseli, Leandro H. M. Oliveira.....366-370

Coleta, composição e etapas de pré-processamento de corpus: procedimentos para a anotação multimodal da FrameNet Brasil

Anna B. C. Silva, Iasmin Rabelo, Igor M. Oliveira, Mariana Souza, Maucha Gamonal, Raquel Roza.....371-375

Abordagens Baseadas em Léxicos para a Classificação de Sentimentos Orientada aos Alvos de Opinião em Comentários do Domínio Político

Lucas Lazarini, Fábio S. Igarashi Anno, Eloize R. Marques Seno, Helena M. Caseli.....376-381

Aryon: um aplicativo Shiny para documentação e análise de línguas indígenas brasileiras

Mateus Zaparoli, Katiuska Rowe, Magnun Rochel Madruga.....382-385

Desambiguação dos termos do Atlas Linguístico do Brasil através da OpenWordnet-PT-ALiB

Augusto Sampaio Barreto, Daniela Barreiro Claro.....386-390

TransAlign: tradução e alinhamento de corpora para a língua portuguesa

Alan Rios Melo, Daniela Barreiro Claro.....391-396

Desafios da tarefa de Extração de Informação Aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual

Beatriz Paixão Queiroz, Rerisson Cavalcante, Daniela Barreiro Claro.....397-401

Técnicas de sumarização de textos jurídicos para suporte à classificação de documentos de decisões judiciais

Hellen Harada, Fabíola Pereira, Alex Almeida, Daniela Freire, Márcio Dias, Nádia Silva, Pedro Andrade, André Carvalho.....402-406

Uso de modelagem de tópicos para agrupamento de notícias: uma abordagem usando BERTopic

Pedro Henrique Pereira, Ticiano Linhares Coelho da Silva.....407-411

Proposta de avaliação da percepção dos impactos da inteligência artificial generativa na educação superior

Ana Luíza Ferreira Vieira, Maria Cecília Zanon De Amorim, Evandro Cunha.....412-416

Explorando Redes Neurais Profundas para Tarefa de Aceitabilidade Linguística

Henrique Santos, Késia P. Alencar, Rogério F. de Sousa, Rafael T. Anchiêta.....417-422

Proceedings of 14th Symposium in Information and Human Language
Technology

STIL 2023

XIV Symposium in Information and Human
Language Technology

Less is More? Investigating Meta-Learning’s Suitability in Sentence Compression for Low-Resource Data

L. Gustavo Coutinho do R.¹, José Antônio F. de Macêdo¹,
Ticiane L. Coelho da Silva¹

¹Insight Data Science Lab – Universidade Federal do Ceará (UFC)

{gustavo.coutinho, jose.macedo, ticianalc}@insightlab.ufc.br

Abstract. *The sentence compression task is essential in the text summarization process. Unfortunately, the lack of labeled data for specific domains restricts the training of deep learning models to address this problem effectively. In this paper, we present an approach using a meta-learning algorithm called MAML to tackle this issue and assess the viability of this technique for the given task, with particular emphasis on its comparison to a fine-tuned BERT model. Our experiments reveal that a simpler approach involving fine-tuning a language model, such as BERT, might indeed be more effective in low-resource scenarios, consistently outperforming the meta-learning techniques for this particular task.*

1. Introduction

The dramatic increase in textual data on the Internet has made it challenging for users to extract valuable information in a reasonable time. Natural Language Processing (NLP) systems can help reduce this workload by performing tasks such as text classification, named entity recognition, and text summarization. In the context of text summarization, sentence compression plays a crucial role in generating concise yet meaningful summaries.

Sentence compression aims to create a shorter version of an input sentence while retaining essential information and ensuring grammatical correctness. There are two types of compressions: extractive and abstractive [Tas and Kiyani 2007]. Extractive compressions remove words without altering the word order, while abstractive compressions generate summaries by rearranging or introducing new words without restrictions.

Existing sentence compression models often rely on large-scale data, but specific domains and Low Resource Languages (LRL) face data scarcity challenges. To address this, we propose employing the few-shot learning paradigm to fine-tune pre-trained models using limited examples from different domains to simulate the low resource scenario. We utilize optimization-based meta-learning techniques, specifically the model-agnostic meta-learning (MAML) algorithm [Finn et al. 2017], to leverage existing models and datasets.

The main challenges in applying meta-learning techniques in sentence compression tasks are:

1. The existence of only one compression for each sentence and one sentence for each compression in the dataset used, unlike image classification tasks that have large datasets, such as Imagenet, and which have several examples for each class, such as cats, dogs, lions, and birds classes.

2. The existence of few datasets for extractive sentence compression, such as the Google News Dataset (GND) that we will use throughout this work, mainly for LRL such as Brazilian Portuguese.
3. Even the most popular dataset for the sentence compression task (GND) does not have any information about grouping the pairs of sentences into categories. Most papers that use meta-learning for few-shot learning tasks group the data for task creation [Mi et al. 2019, Yu et al. 2018].

To the best of authors’ knowledge, no paper addresses these problems for the sentence compression task.

The main contributions of this paper include the following:

- We present a novel approach to addressing the sentence compression task by framing it as a Named Entity Recognition (NER) problem.
- We propose a method for modeling the sentence compression task as a meta-learning problem, utilizing few-shot learning principles and the well-established model-agnostic meta-learning (MAML) algorithm.
- We conduct comprehensive evaluations using the Google News dataset, exploring various few-shot scenarios, dataset divisions, and comparing with a BERT-based solution.

Our findings indicate that while models developed using the meta-learning approach can rapidly adapt to new tasks, they do not surpass the performance of a BERT model fine-tuned with the same dataset. This observation somewhat contrasts with findings in the context of other NLP tasks [Mi et al. 2019, Yu et al. 2018].

2. Related Works

Traditional Methodologies for Sentence Compression: Early approaches to extractive sentence compression relied on parsing tree pruning methods [Filippova and Altun 2013]. However, these techniques were prone to errors in constructing the trees themselves. Contemporary strategies began to frame the problem as Seq2Seq tasks, incorporating various modifications in LSTM-based models [Filippova et al. 2015, Soares et al. 2020, Kamigaito and Okumura 2020]. To the best of the author’s knowledge, none of these methods effectively address the challenge posed by the limited availability of training data.

Meta-Learning in NLP Tasks: Meta-learning techniques have gained traction in various Natural Language Processing tasks [Lee et al. 2022]. One reason for the growing popularity of meta-learning is its effectiveness in low-resource situations, where collecting and annotating datasets can be prohibitively expensive, such as in Natural Language Generation [Mi et al. 2019, Qian and Yu 2019] and Machine Translation [Gu et al. 2018] tasks. Another factor driving the adoption of these techniques is the domain shift between training data and real-world testing and application scenarios [Li et al. 2020, Song et al. 2019]. The sentence compression problem addressed in this study faces both of these challenges.

3. Meta-Learning for Low-Resource Sentence Compression

One of the goals of this study is to develop an efficient model for the sentence compression task in low-resource scenarios, such as those found in medical or law enforcement

domains. To achieve this, we propose framing the task as a meta-learning problem, enabling the model to leverage knowledge from related tasks. The modeling process consists of four main steps:

1. Reformulate the sentence compression problem as a NER problem, utilizing a pre-trained BERT model as the foundation,
2. Create a dataset comprised of many sentence compression tasks,
3. Train a meta-model using the created tasks and the base BERT model,
4. Perform adaptation of the meta-model for a new sentence compression task, potentially within a distinct domain.

The following subsections provide a detailed description of each of these steps in the proposed methodology.

3.1. Sentence Compression as a NER problem

The NER problem is a well-established task in NLP, which involves identifying and classifying named entities within a given text. The goal of redefining sentence compression as an NER problem is to take advantage of the vast knowledge and tools available for NER and use them for the sentence compression task. By doing so, we aim to exploit the pre-existing strengths of BERT models in capturing contextual information and handling diverse linguistic structures for improved compression performance [Ma et al. 2019].

The process of task adaptation consists of assigning an entity label to each word in the dataset’s compressions. We use two entity labels: `keep` for words that should be retained in the compressed sentence and `compress` for words that should be omitted. This transformation allows us to treat sentence compression as a NER task, where the objective is to identify and classify words in a sentence as either `keep` or `compress`.

To train a model to solve this adapted task, we employ a pre-trained BERT model as the foundation, given its proven success in various NLP tasks, including NER. We add an additional classification layer to the model to classify each word of each sentence as either `keep` or `compress`. The model is then fine-tuned on the newly created NER-style sentence compression dataset, learning to recognize and classify essential words and phrases to be retained in the compressed sentence.

3.2. Dataset creation

When modeling a problem using the meta-learning approach, we assume access to a distribution of tasks $\mathcal{P}(\mathcal{T})$. The objective of assuming this distribution is to sample T_i tasks from the $\mathcal{P}(\mathcal{T})$ distribution – where T_i is a task composed of a training set ($D_{T_i}^{Train}$, also known as support set) and a test set ($D_{T_i}^{Test}$, also known as query set) – to train a meta-model that can generalize well to all tasks used in the training process. The trained meta-model can then be fine-tuned to a task T' , also sampled from the $\mathcal{P}(\mathcal{T})$ distribution, that was not seen in the meta-model training [Bansal et al. 2021]. When dealing with supervised learning tasks using meta-learning, we create the $\mathcal{P}(\mathcal{T})$ distribution based on a fixed set of tasks, subsampled from all classes [Vinyals et al. 2016].

In our sentence compression tasks, a sentence is an input x and its compression is an output y . An example task includes a Support Set (or D_{Train}) of five sentences with their compressions, and a Query Set (or D_{Test}) of five distinct sentences and their corresponding compressions from the Support Set. Mapped to the N -way, K -shot learning scenario, it yields values of $N = 5$ and $K = 1$.

3.3. Meta-training a model

Following the creation of the dataset, we utilized the Model-Agnostic Meta-Learning (MAML) algorithm to train a meta-model. This model is designed to process a range of sentence compression tasks as input. The meta-model we chose is an instance of the BERT model, which consists of approximately 107 million parameters.

The meta-model will be initialized with random θ^{Meta} parameters (also called meta-parameters). In an iteration step of the MAML algorithm, each task T_i will use a copy of all the parameters θ_i (here called task parameters) of the meta-model and will optimize them using the dataset $D_{T_i}^{Train}$, generating updated parameters θ'_i . Once all tasks are optimized, the datasets $D_{T_i}^{Test}$ will be used to optimize the meta-parameters θ^{Meta} using the parameters θ'_i of the tasks that were calculated. This way, the meta-model will be generalizing all tasks used as input in its own meta-parameters.

3.4. Fine-tuning process

With the meta-model trained from N -way, K -shot learning tasks, we obtain the parameters θ^{Meta} that should generalize the tasks used in the previous step. Given a new sentence compression task with a small dataset for training, we can use the newly trained meta-model for the fine-tuning process. The θ^{Meta} parameters will be used to initialize a new model and update it with the data from the new task. Since the new task has a small amount of associated data, it will take advantage of the already trained parameters of the meta-model as a good starting point for optimization.

4. Experiments

In this section, we will first formally present the Research Questions (RQ) that this work aims to answer:

- RQ1. How does a model trained using meta-learning and conventional machine learning compare?
- RQ2. What is the change in the performance of a meta-model if we increase the number of examples per class of each class?

In summary, the RQs presented try to evaluate if using meta-learning techniques is viable for the sentence compression problem.

We will also present the baseline and experiment settings, the dataset used and the data augmentation process applied, and finally, the results of the experiments performed.

4.1. Baseline and Model Settings

Since MAML is model agnostic, for all experiments performed in this work, we used as a baseline for the conventional machine learning training a BERT model [Kenton and Toutanova 2019] with an additional classification layer to classify the words with the entities `keep` or `compress`. For the MAML implementation, we used the Pytorch framework.

For all experiments performed, we considered two different settings:

- Scratch-BERTSC: Fine tune the BERT with only the target N -way, K -shot low-resource task.

- Meta-BERTSC: Train a meta-model using MAML and different N -way, K -shot low-resource tasks and then fine-tune it to an unseen task.

In the Meta-BERTSC setting, the number of tasks used to train the meta-model, i.e., tasks batch size, is one of the parameters that will be evaluated in the experiments, varying among 16, 32, 64 and 128. We set $\alpha = 0.001$ and $\beta = 0.0001$. A last parameter evaluated in this experiment is the size of each task: we consider a 5-way, 1-shot scenario and a 5-way, 5-shot scenario.

All models, for both Scratch-BERTSC and META-BERTSC, were evaluated in the target task. For the evaluation metrics, we used the ROUGE score variances, i.e., ROUGE-1, ROUGE-2 and ROUGE-L. We will also present the accuracy for the trained models. To assess the similarity between the compression ratio of system outputs and that of gold compressed sentences, we employed the delta compression (ΔC) metric, which is the difference between the system compression ratio and gold compression ratio [Kamigaito et al. 2018].

In terms of training times, the Scratch-BERTSC model outperformed the META-BERTSC approach. For a batch size of 16 tasks in a 5-way, 1-shot scenario, for example, Scratch-BERTSC completed training in 49 seconds, while META-BERTSC took 168 seconds. Similarly, for a larger batch size of 128 tasks, Scratch-BERTSC required 366 seconds, and META-BERTSC required considerably more time, 1326 seconds.

4.2. Google News Dataset for SC

The Google News Dataset [Filippova and Altun 2013], comprising 200,000 pairs of news article headlines and their compressed versions, is used for all experiments in this work. Despite the categorized nature of news articles, this dataset doesn't provide information about the pairs' categories.

We adapted the Google News dataset into a meta-learning framework using the pipeline from Subsection 3.2, dividing it randomly into 5-way, 1-shot learning tasks. This implies each task has examples from five different classes, with one example per class, considering each compression as a class with one sentence generating that specific compression.

Unfortunately, with the original Google News Dataset, it is not possible to vary the size of K in an N -way, K -shot learning sentence compression task since it is practically impossible for two or more non-related sentences to have the same compression, i.e., two or more examples with the same class. We used data augmentation techniques to work around this problem to create more sentences based on one single compression and evaluate the scenario of increasing the number of N and K in the meta-model training.

4.3. Compression Augmentation

For the compression augmentation process, we used a pre-trained BERT model [Kenton and Toutanova 2019] on English language with a masked language modeling (MLM) objective¹.

For each compression of our dataset, we randomly place mask tokens between the words. We empirically choose to insert 10% of the number of words as masks in the

¹The pre-trained model used is available at <https://huggingface.co/bert-base-cased>

compression (or one mask if the compression had less than ten words). With too many masks, the sentence generated might not be semantically correct. Once all masks are placed, we use the previously mentioned pre-trained model to replace them with actual correct words.

This whole process can be repeated to generate any number of sentences based on a single compression. Figure 1 presents an example of the compression augmentation process.

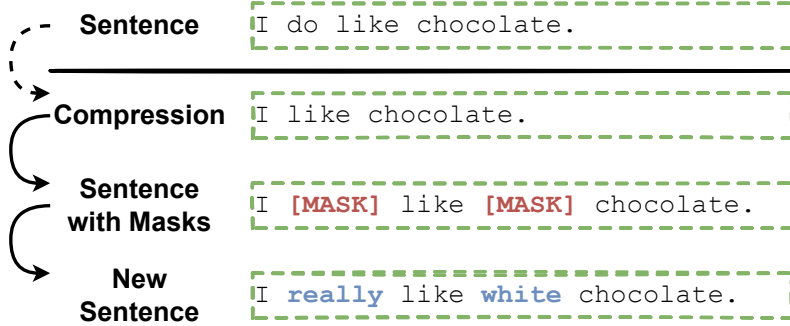


Figure 1. Example of a compression augmentation using a masked language modeling (MLM) objective.

After generating new sentences based on each compression, we are now able to create N -way, K -shot learning tasks with larger values of K : the task will have more than one sentence leading to the same compression (to do the reverse process of data augmentation, i.e., compress the sentence, we only need to extract the words added by the model to get the respective compression).

To verify that the generated sentences have the same meaning as the compressions that originated them, the sentence-level embeddings model Universal Sentence Encoder [Cer et al. 2018] were calculated for all sentences and compressions, and the values of each pair were calculated using the cosine similarity. We calculated the embedding similarities and extracted some descriptive statistics for the set of all comparisons: mean=0.95440, median=0.96100, and standard deviation=0.03443. These values show that the meaning of the newly generated sentence is very similar to the meaning of the original compression.

4.4. Results and Discussion

Meta-learning Versus Conventional Learning: To answer the RQ1., we compared two models created using the META-BERTSC and the Scratch-BERTSC settings described previously.

We trained the model from the META-BERTSC setting with 5-way, 1-shot randomly created tasks, i.e., we randomly chose five pairs of sentences and compression from the Google News Dataset to create each task. We vary the batch size of tasks from the list of values [16, 32, 64, 128].

The rest of the dataset was used to sample sentence compression tasks to fine-tune the meta-model chosen previously individually. Table 1 presents the average F_1 Score

(F_1), ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) and ΔC metrics of all models fine-tuned for [16, 32, 64, 128] tasks.

batch	F_1	R-1	R-2	R-L	ΔC
16	0.78002	0.76478	0.52067	0.76478	0.10163
32	0.77046	0.74560	0.50840	0.74522	0.08339
64	0.74938	0.72042	0.47680	0.72006	0.07716
128	0.78230	0.75714	0.52604	0.75606	0.08180

Table 1. Results for the 5-way, 1-shot learning scenario with the model trained from the META-BERTSC setting.

Table 2 presents the results of the model trained in the Scratch-BERTSC scenario with the same tasks and metrics used in the fine-tuning process for the META-BERTSC scenario described previously.

batch	F_1	R-1	R-2	R-L	ΔC
16	0.80777	0.77974	0.60439	0.77860	0.047612
32	0.80951	0.76883	0.59564	0.76835	0.046216
64	0.79398	0.75405	0.57421	0.75345	0.050612
128	0.80915	0.76810	0.59827	0.76682	0.041529

Table 2. Results for the 5-way, 1-shot learning scenario with the model trained from the Scratch-BERTSC setting.

The fine-tuned model Scratch-BERTSC consistently outperforms the META-BERTSC in all different metrics considered. Table 3 shows two examples of predicted compressions with the Scratch-BERTSC approach and the META-BERTSC approach.

Sentence:	Asda has dropped Saatchi & Saatchi out of the pitch for its £100m advertising account, ending its 20-year relationship with owner Publicis Groupe.
Compression:	Asda has dropped Saatchi & Saatchi for its £ 100m account.
[S-BSC]:	Asda has dropped Saatchi & Saatchi for its £100m account.
[M-BSC]:	Asda has dropped Saatchi Saatchi accounts for its £100m account.

Table 3. An example of compressions performed by the two models evaluated. In this table, the terms “S-BSC” and “M-BSC” are abbreviations for the terms “Scratch-BERTSC Prediction” and “META-BERTSC Prediction”, respectively.

Increasing the Number K of Examples per Class: To answer RQ2., we compared two models created using the META-BERTSC setting: one with 5-way, 1-shot randomly created tasks (results from Table 1) and another with 5-way, 5-shot randomly created tasks. We also fine tuned the Scratch-BERTSC model with the same increased amount of tasks for the 5-way, 5-shot setting.

Table 4 and Table 5 presents the average F_1 Score (F_1), ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) and ΔC metrics of all models fine-tuned for [16, 32, 64, 128] tasks.

batch	F_1	R-1	R-2	R-L	ΔC
16	0.75577	0.72926	0.49441	0.72828	0.06936
32	0.75175	0.71070	0.46656	0.70835	0.03720
64	0.75616	0.72122	0.48416	0.71966	0.05984
128	0.76329	0.73454	0.50208	0.73355	0.06201

Table 4. Results for the 5-way, 5-shot learning scenario with the model trained from the META-BERTSC setting.

batch	F_1	R-1	R-2	R-L	ΔC
16	0.80318	0.75739	0.59899	0.75519	0.02642
32	0.80239	0.76228	0.60694	0.75975	0.02743
64	0.79808	0.74732	0.57472	0.74573	0.03689
128	0.76638	0.71441	0.49862	0.71321	0.02050

Table 5. Results for the 5-way, 5-shot learning scenario with the model trained from the Scratch-BERTSC setting.

Based on the presented results, we cannot guarantee that a higher value of K generate better results, since for none of the two approaches are better than the other for all metrics, both for the Scratch-BERTSC and META-BERTSC approaches. Unfortunately there is a current limitation in the increase of the value of K because it would demand the generation of more sentences based on the same compression. The automatic addition of too many new words in the compression could generate syntactically incorrect sentences.

5. Conclusions and Future Works

In this work, we used the MAML algorithm for sentence compression via meta-learning, particularly when data is scarce. The resulting meta-model could be fine-tuned for specific domains with little data. However, our findings revealed its limitations, as it didn’t outperform a fine-tuned BERT model, even with limited data. This suggests BERT’s robustness and contextual understanding, combined with fine-tuning, may be a better approach for sentence compression in low-resource scenarios.

In future works, we intend to investigate the following additional research questions regarding the similarity between the distribution of tasks and the meaning bias between a sentence and the compressions generated during the work:

- RQ1 What would be the impact of training a meta-model with domain-specific tasks, e.g., sentence compression task with sentences only related to sports, and fine-tune it to a different domain task?
- RQ2 How similar are the tasks of the distribution used? And when we separate by domain, how similar are the tasks? And the new tasks that the meta-model will be tuned for, how similar are they to each other?

Finally, the present work can also serve as a basis for researchers who wish to investigate the advantages of leveraging pre-trained BERT models and fine-tuning techniques and the use of meta-learning in the sentence compression task since we believe this is the first study to propose this approach to this problem.

References

- Bansal, T., Gunasekaran, K. P., Wang, T., Munkhdalai, T., and McCallum, A. (2021). Diverse distributions of self-supervised tasks for meta-learning in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5812–5824, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal. Association for Computational Linguistics.
- Filippova, K. and Altun, Y. (2013). Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.
- Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135. PMLR.
- Gu, J., Wang, Y., Chen, Y., Li, V. O. K., and Cho, K. (2018). Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.
- Kamigaito, H., Hayashi, K., Hirao, T., and Nagata, M. (2018). Higher-order syntactic attention network for longer sentence compression. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1716–1726, New Orleans, Louisiana. Association for Computational Linguistics.
- Kamigaito, H. and Okumura, M. (2020). Syntactically look-ahead attention network for sentence compression. In *Proceedings of the AAAI*, volume 34, pages 8050–8057.
- Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Lee, H.-y., Li, S.-W., and Vu, T. (2022). Meta learning for natural language processing: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 666–684, Seattle, United States. Association for Computational Linguistics.
- Li, J., Shang, S., and Shao, L. (2020). Metaner: Named entity recognition with meta-learning. In *Proceedings of The Web Conference 2020*, pages 429–440.
- Ma, X., Xu, P., Wang, Z., Nallapati, R., and Xiang, B. (2019). Domain adaptation with bert-based domain classification and data selection. In *Proceedings of the 2nd DeepLo*, pages 76–83.

- Mi, F., Huang, M., Zhang, J., and Faltings, B. (2019). Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the 28th IJCAI*, pages 3151–3157.
- Qian, K. and Yu, Z. (2019). Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy. Association for Computational Linguistics.
- Soares, F. M., da Silva, T. L. C., and de Macêdo, J. F. (2020). Sentence compression on domains with restricted labeled data. In *Proceedings of the 12th ICAART*, pages 130–140.
- Song, Y., Liu, Z., Bi, W., Yan, R., and Zhang, M. (2019). Learning to customize language model for generation-based dialog systems. *CoRR*, abs/1910.14326.
- Tas, O. and Kiyani, F. (2007). A survey automatic text summarization. *PressAcademia Procedia*, 5(1):205–213.
- Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *NeurIPS*, 29.
- Yu, M., Guo, X., Yi, J., Chang, S., Potdar, S., Cheng, Y., Tesauro, G., Wang, H., and Zhou, B. (2018). Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215, New Orleans, Louisiana. Association for Computational Linguistics.

Predição de transtorno depressivo em redes sociais: BERT supervisionado ou ChatGPT *zero-shot*?

Wesley Ramos dos Santos¹ ,
Ivandr  Paraboni¹

¹University of S o Paulo (EACH-USP)
Av Arlindo Bettio 1000, S o Paulo, Brazil

{wesley.ramos.santos, ivandre}@usp.br

Abstract. *This article presents a first study on the use of the ChatGPT dialogue system in a complex and sensitive application, namely, the computational prediction of mental health disorders from social media text. To this end, we conducted an experiment to compare a traditional supervised approach based on BERT with a zero-shot strategy based on natural language prompts submitted directly to the dialogue system. Results of this evaluation, taking into account the accuracy of the classification task in view of the need for corpus annotation in the supervised approach, highlight different advantages of each alternative.*

Resumo. *Este artigo apresenta um primeiro estudo sobre o uso do sistema de di logo ChatGPT em uma aplica  o complexa e sens vel: a predi  o computacional de transtornos de sa de mental a partir de textos provenientes de redes sociais. Para esse fim, foi conduzido um experimento comparando uma abordagem supervisionada tradicional baseada em BERT com uma estrat gia zero-shot baseada em prompts em l ngua natural submetidos diretamente ao sistema de di logo. Resultados desta avalia  o, levando em conta a acur cia da tarefa de classifica  o face   necessidade de anota  o pr via de c rpus da abordagem supervisionada, destacam diferentes vantagens de cada alternativa.*

1. Introdu  o

As formas de representa  o textual utilizadas em aplica  es de PLN t m evolu do de forma acelerada em anos recentes. Partindo-se de modelos simples baseados em contagens de *tokens*, como o *bag-of-words* e suas varia  es, no espa o de poucos anos passou-se ao uso de *word embeddings* est ticos, como os produzidos com word2vec [Mikolov et al. 2013], e posteriormente dependentes de contexto, como nos modelos de l ngua pr -treinados do tipo BERT [Devlin et al. 2019]. Este  ltimo, at  recentemente considerado o modelo mais expressivo de sua categoria, com 340 milh es de par metros, come a a enfrentar a concorr ncia de modelos ainda maiores e mais sofisticados, denominados LLMs (*large language models*) como por exemplo GPT-3 [Brown et al. 2020] e Bloom [BigScience Workshop 2022], com 175 bilh es de par metros cada.

A crescente complexidade dos atuais LLMs pode entretanto restringir seu uso a aplica  es para o idioma ingl s, ou  quelas em que   poss vel fazer uso de uma vers o multil ngue do modelo. Neste sentido, uma alternativa de interesse para a experimenta  o

em PLN foi a disponibilização da interface de diálogo ChatGPT¹, que combina um LLM da família GPT com métodos de aprendizado supervisionado e por reforço utilizando *feedback* humano para modelar a tarefa de conversação com usuários humanos com alto grau de realismo.

Embora sejam modelos do tipo generativo (i.e., que essencialmente completam um trecho de texto com sua continuação mais provável), LLMs e sistemas deles derivados (como ChatGPT) possuem, em virtude do grande volume de dados de treinamento a que tiveram acesso durante sua construção, ampla capacidade de associar pares de textos, e podem assim ser facilmente adaptados a diversas tarefas de interpretação de língua natural [Zhang et al. 2023]. Em especial, observa-se que com poucos meses de lançamento o sistema ChatGPT já começou a ser cogitado como uma possível alternativa a métodos supervisionados tradicionais (i.e., baseados em *córpus* de exemplos rotulados) por permitir a consulta direta ao LLM sem exemplos prévios. Em métodos deste tipo, ao invés de rotular um *córpus* de avaliações de produtos com informação de sentimento (e.g., positivo, neutro ou negativo), podemos em tese simplesmente submeter ao LLM uma instrução em língua natural perguntando qual o sentimento expresso em um determinado texto.

A ausência de exemplos de treinamento em métodos baseados em LLMs é uma forma de classificação de texto do tipo *zero-shot*. Embora haja considerações relativas à segurança e contaminação de dados (i.e., o fato de que estes modelos são constantemente atualizados e podem ter sido expostos aos dados de teste da tarefa em tempo de treinamento, cf. [Zhang et al. 2023]), o uso de ferramentas como ChatGPT em tarefas de PLN como sumarização, sistemas de respostas a perguntas, análise de sentimentos e outras [Qin et al. 2023] tem se difundido com rapidez, e recentemente foi levantada até mesmo a hipótese do ‘começo do fim da tarefa de anotação de *córpus*’ [Kuzman et al. 2023].

Com base nestas considerações, neste trabalho apresentamos um primeiro estudo sobre o uso de ChatGPT em uma aplicação complexa e enfocando um tema intencionalmente sensível, carregado de questões éticas do tipo que sistemas como ChatGPT notoriamente tentam evitar: a predição computacional de transtornos de saúde mental a partir de textos provenientes de redes sociais. Aplicações deste tipo, já amplamente desenvolvidas com uso de métodos supervisionados convencionais [Chancellor and Choudhury 2020, Su et al. 2020], são aqui tratadas pela primeira vez com uso de métodos *zero-shot*, não havendo (até onde temos conhecimento) risco de contaminação de dados.

De forma mais específica, o presente trabalho objetiva avaliar a detecção de indivíduos com maior risco de desenvolver transtorno depressivo a partir de suas postagens no Twitter brasileiro, utilizando para este fim uma abordagem supervisionada tradicional baseada em BERT e, como alternativa, uma estratégia do tipo *zero-shot* baseada em *prompts* submetidos diretamente ao sistema de diálogo ChatGPT, levando-se em conta a acurácia da tarefa de classificação face à necessidade de anotação prévia de *córpus* da abordagem supervisionada. As principais contribuições previstas são as seguintes:

- Método inédito do tipo *zero-shot* para detecção de transtorno depressivo a partir de postagens no Twitter em Português.
- Comparação com modelo supervisionado do tipo estado-da-arte para essa tarefa, baseado em BERT e Bi-LSTMs.

¹<https://chat.openai.com/chat>

O restante deste artigo está organizado da seguinte forma. A seção 2 apresenta um breve levantamento de estudos existentes da área de predição de transtorno de depressão a partir de textos. A seção 3 introduz os modelos baseados em BERT e ChatGPT desenvolvidos. A seção 4 descreve a avaliação conduzida, e a seção 5 apresenta seus resultados. Finalmente, a seção 6 sumariza a presente discussão e apresenta futuras direções de pesquisa sobre o assunto.

Considerações éticas

Os modelos computacionais discutidos foram desenvolvidos com base em dados publicamente disponibilizados na plataforma Twitter, aqui tratados de forma anonimizada e confidencial. A presente abordagem linguístico-computacional não deve ser vista como substituto a outras formas de aquisição de conhecimento (em especial, derivadas da área médica) e não objetiva diagnosticar *indivíduos* com transtornos de saúde mental, mas apenas contribuir para a área de análise computacional de redes sociais enfocando o *estudo da linguagem* empregada nestas circunstâncias.

2. Trabalhos relacionados

A detecção transtorno depressivo com base em dados textuais (e.g., provenientes de redes sociais ou outras fontes) é tipicamente modelada na forma de um problema de aprendizado de máquina supervisionado, ou seja, fazendo uso de *corpus* de textos rotuladas com informações relativas ao estado de saúde mental de seus autores (e.g., usuários de redes sociais) para treino e teste de classificadores. Sob esta perspectiva, a tarefa pode ser vista como uma instância do problema de caracterização autorral [dos Santos et al. 2020b, Pavan et al. 2023, Flores et al. 2022] combinado à detecção de linguagem afetiva [da Silva et al. 2020]. Um levantamento de estudos recentes deste tipo é apresentado na Tabela 1, com indicação do gênero de texto considerado (Reddit, Twitter), a forma de representação dos dados textuais (b=*bag of words*, BERT [Devlin et al. 2019], d=características de domínio, e=*embeddings*, h=horário da publicação, i=imagens, l=atributos LIWC [Pennebaker et al. 2001], m=metadados, n=informações de rede, p=*part-of-speech*, s=atributos afetivos, t=tópicos, u=informações demográficas), e métodos computacionais (e.g., CNN=redes neurais convolucionais, LSTM=*long short-term neural networks*, LR=regressão logística, RF=*Random Forest*, etc.).

Dentre os estudos selecionados, observa-se uma ligeira predominância de trabalhos baseados na rede social Reddit. Esta preferência pode ser explicada pela maior facilidade de acesso e reúso de dados desse tipo para fins de pesquisa, o que é mais restrito no caso da plataforma Twitter. Assim, postagens Reddit são usadas em alguns dos conjuntos de dados mais conhecidos para o idioma inglês, como os *corpus* SMHD [Cohan et al. 2018] e eRisk [Losada and Crestani 2016], sendo este último também a base de uma série de desafios computacionais (ou ‘*shared tasks*’) *Early Risk Prediction on the Internet* [Parapar et al. 2022].

Quanto aos tipos de modelos textuais utilizados, a Tabela 1 reflete a evolução natural da pesquisa em áreas correlatas do PLN, com predominância inicial de modelos do tipo *bag-of-words* e engenharias de características, e sua substituição gradual por modelos baseados em *word embeddings* e, mais recentemente, BERT [Devlin et al. 2019].

Tabela 1. Detecção de transtorno depressivo a partir de texto

Estudo	Gênero	Repres. textual	Método
[Cohan et al. 2018]	reddit	b,e	FastText
[Trotzek et al. 2018]	reddit	e,p,m,d	CNN
[Kumar et al. 2019]	twitter	d,s,h	ensemble
[Aragón et al. 2019]	reddit	s	SVM
[Cacheda et al. 2019]	reddit	h,m,n	RF
[Burdisso et al. 2020]	reddit	b	SS3
[Lin et al. 2020]	twitter	e,i	CNN
[Yazdavar et al. 2020]	twitter	b,s,t,i,n,l,u	RF
[Souza et al. 2020]	reddit	e	LSTM
[Souza et al. 2021]	reddit	e	LSTM+CNN
[Ansari and Ji 2022]	reddit,twitter	e,s	LR+LSTM
[dos Santos et al. 2023]	twitter	BERT	Bi-LSTM

No que diz respeito aos métodos computacionais empregados, de modo geral observa-se a mesma trajetória, com o uso de classificadores tradicionais baseados em contagens de *tokens* sendo gradualmente substituído por métodos de classificação de sequências baseados em aprendizado profundo, incluindo o uso mais recente de arquiteturas baseadas em *transformers*.

Com exceção dos estudos para o português em [dos Santos et al. 2020a, dos Santos et al. 2023], que introduziram o *cópus* denominado SetembroBR para detecção de transtorno de depressão e ansiedade no Twitter brasileiro, todos os trabalhos identificados são dedicados ao idioma inglês. Assim, este *cópus* será tomado como base no presente trabalho, conforme discutido nas próximas seções.

Finalmente, observa-se que nenhum dos estudos identificados faz uso de métodos *zero-shot*, baseados em *prompt* ou em modelos de língua de grande escala e similares. Destacamos, entretanto, que o sistema ChatGPT tem sido utilizado da forma que pretendemos aplicar ao presente problema de detecção de depressão em diversas outras tarefas de PLN, como a classificação de gênero linguístico [Kuzman et al. 2023], detecção de atitudes ou posicionamentos [Zhang et al. 2023] e análise de sentimentos [Qin et al. 2023].

3. Modelos de predição de transtorno depressivo

O estudo conduzido avalia dois modelos de predição de transtorno depressivo a partir de textos publicados na rede social Twitter. O primeiro modelo, baseado no sistema de diálogos ChatGPT, é uma abordagem do tipo *zero-shot* que não faz uso de nenhum tipo de anotação manual de dados. O segundo modelo segue uma abordagem supervisionada tradicional baseada em dados de um *cópus* rotulado e uso do modelo de língua BERT [Devlin et al. 2019]. O objetivo da avaliação é o de verificar qual destas duas abordagens apresenta melhores resultados para um subconjunto de *timelines* (conjuntos de postagens ou *tweets*) de teste do *cópus* SetembroBR [dos Santos et al. 2023].

O modelo de predição baseado no sistema ChatGPT foi construído da seguinte forma. Quando questionado diretamente sobre saúde mental ou outras questões de natureza sensível, o sistema ChatGPT geralmente evita fornecer uma resposta, usando justifi-

cativas como ‘o diagnóstico de depressão só pode ser feito por um médico’, ‘sou apenas um modelo de língua’ e similares. Assim, como forma de contornar estas medidas de proteção, foi utilizado um *prompt* que solicita informações sobre possíveis indicadores linguísticos de depressão que podem ou não estar presentes em um dado *tweet*. O *prompt* submetido ao sistema ChatGPT 14 March Version foi o seguinte:

Desconsidere o diálogo anterior e avalie se o texto a seguir apresenta indicadores linguísticos de depressão em uma escala de 0 a 10, onde 0 significa nenhum indício e 10 significa indícios muito fortes. Ao responder, informe apenas um número de 0 a 10, sem justificar ou explicar sua resposta. O texto a ser analisado é <texto do tweet aqui>.

A escala de 0 a 10 foi utilizada porque, em testes preliminares, este tipo de *prompt* se revelou mais preciso do que solicitações mais diretas, como no caso de resposta binária (e.g., ‘indique se o texto possui ou não indicadores de depressão’). Além disso, sendo um sistema de conversação dotado de memória de diálogo, foi solicitado explicitamente que as interações anteriores (que podiam ser referentes ao mesmo autor do *tweet* atual ou outro qualquer) fossem desconsideradas, forçando o sistema a analisar cada mensagem de forma independente das demais. Finalmente, para maior rapidez na resposta do sistema, foi solicitado que não fosse apresentada nenhuma explicação adicional motivando a resposta.

Com base neste método, uma coleção de *tweets* de teste foi rotulada pelo sistema ChatGPT com escores de 0 a 10. Para a combinação desses escores individuais em um rótulo de classe global (i.e., considerando todos os *tweets* de uma *timeline* de um determinado indivíduo), a média destes escores foi comparada a um valor de *threshold* fixo previamente computado a partir de dados de treino não utilizados na presente avaliação.

De forma mais específica, o sistema ChatGPT foi utilizado para rotular um conjunto de 30 *timelines* de treino da classe de *Controle* (ou seja, um conjunto de indivíduos aleatórios selecionados a partir da população geral) contendo 80 *tweets* cada. A seguir, foi computada a média de escores de todos os $30 * 80 = 2400$ *tweets* de treino, e este valor foi utilizado como *threshold* para definir os rótulos (*Diagnosticados* ou *Controle*) das *timelines* de teste.

Como alternativa ao modelo baseado em ChatGPT, foi considerada também uma abordagem tradicional de aprendizado supervisionado baseada em modelos de língua do tipo BERT. Para este fim, utilizou-se o modelo BERTabaporu [da Costa et al. 2023], um modelo BERT treinado com base em 2.9 bilhões de *tokens* obtidos a partir de 237 milhões de *tweets* em português. Nesta abordagem, a representação das *timelines* a serem rotuladas como *Diagnosticado* ou *Controle* é feita em sequências consecutivas de 10 *tweets* iniciadas em uma posição aleatória da *timeline* a cada época. Esta representação textual alimenta uma camada Bi-LSTM com 100 neurônios seguida de 3 camadas do tipo MLP, cada uma com dropout de 0,1 e função de ativação *softmax*.

4. Avaliação

Como forma de comparar o desempenho dos modelos baseados em ChatGPT e BERT discutidos na seção anterior, foi conduzido um experimento de aplicação destes modelos preditivos a uma porção do corpus SetembroBR [dos Santos et al. 2023] de *timelines* de usuários do Twitter brasileiro com diagnóstico de depressão, e de usuários

Tabela 2. Subconjunto Depressão do cópús SetembroBR

Métrica	Diagnosticados	Controle
Usuários	1684	11788
<i>Tweets</i> (milhões)	2,43	16,99
<i>Tokens</i> (milhões)	29,32	201,94

aleatórios formando um grupo de controle de proporção 7 vezes superior. Nesta definição do problema, também seguida em [Coppersmith et al. 2015, Losada et al. 2017, Lynn et al. 2018, Cohan et al. 2018, Losada et al. 2019, Parapar et al. 2022] e outros, o objetivo é distinguir indivíduos depressivos da população em geral, e não distinguir indivíduos depressivos de não-depressivos.

A Tabela 2 sumariza estatísticas descritivas da porção de dados referentes ao trans-torno depressivo presentes no cópús.

O cópús possui uma divisão aleatória pré-definida entre *timelines* de treinamento (80%) e teste (20%) que foi respeitada no experimento realizado. Entretanto, os dados de treino são usados de forma diferente pelos dois modelos desenvolvidos. Para o modelo baseado em ChatGPT, a porção de treino foi utilizada apenas para cálculo do valor de *threshold* de separação das classes *Diagnosticados* e *Controle*, conforme descrito na seção anterior. O modelo BERT, por outro lado, utiliza a porção de treino completa do cópús para sua construção.

No que diz respeito aos dados de teste utilizados por ambos os modelos, observa-se que não seria praticável submeter manualmente para avaliação do sistema ChatGPT cada um dos cerca de 3,9 milhões de *tweets* de teste do cópús SetembroBR. Assim, na presente avaliação foi utilizado apenas um subconjunto reduzido de *timelines*, contendo cada uma um número também reduzido de *tweets*.

O experimento realizado baseou-se em um subconjunto de 50 *timelines* de cada classe, selecionadas aleatoriamente a partir do conjunto de teste do cópús, e cobrindo cada uma um intervalo fixo de 80 *tweets* consecutivos com a maior frequência possível dos termos ‘depressão’ e ‘ansiedade’. Esta estratégia de seleção objetivou maximizar as chances de que, mesmo analisando-se uma porção reduzida dos dados, algum indício de discussão sobre questões de saúde mental pudesse ser encontrado no trecho avaliado, ressaltando-se que esta simplificação não representa uma vantagem para nenhum dos dois modelos sob avaliação (já que ambos utilizam os mesmos dados de teste), e não torna a tarefa computacional menos complexa (dado que tanto indivíduos das classe *Diagnosticados* como *Controle* podem mencionar ou não estes termos).

5. Resultados

Os dados de teste selecionados foram submetidos para avaliação dos modelos ChatGPT e BERT conforme descrito nas seções anteriores. Para este fim, foram computadas as medidas de acurácia por classe e a acurácia média de cada modelo, observando-se que o conjunto de teste é perfeitamente balanceado. A Tabela 3 apresenta os resultados obtidos.

Os presentes resultados motivam uma série de considerações. Em primeiro lugar, observa-se que, na média global, os dois modelos são essencialmente similares, um resultado que é em certo sentido inesperado tendo-se em vista a grande diferença metodológica

Tabela 3. Resultados de predição de transtorno de depressão.

Modelo	Diagnosticados	Controle	Média
ChatGPT	0,70	0,60	0,65
BERT	0,48	0,84	0,66

entre as duas abordagens. Entretanto, observando-se o comportamento individual de cada modelo nas classes *Diagnosticados* e *Controle*, esta diferença se reflete de forma mais evidente. O modelo baseado em ChatGPT possui capacidade relativamente elevada de classificar corretamente as *timelines* de indivíduos *Diagnosticados*, mas apresenta menor sucesso ao tratar o grupo *Controle*. No caso do modelo baseado em BERT, o efeito é o contrário, ou seja, uma menor acurácia na classe *Diagnosticados* é compensada por uma melhoria na classe *Controle*.

Uma possível explicação para estes resultados seria a de que o modelo baseado em ChatGPT é realmente superior ao modelo baseado em BERT quando aplicado à tarefa de identificar indicadores linguísticos de depressão mas, como a tarefa modelada pelo *cópus* é a de distinção entre *Diagnosticados* e um grupo de *Controle* aleatório (e que não representa uma classe negativa do tipo ‘não diagnosticados’, mas apenas uma população média), o conjunto *Controle* também apresenta alguns indicadores deste tipo (ainda que certamente em menor proporção do que no conjunto *Diagnosticados*). Assim, embora o modelo ChatGPT tenha ampla vantagem em relação ao modelo BERT na identificação destes indicadores, a separação entre estes casos e os exemplos aleatórios carece da noção de ‘população média’ que, crucialmente, está presente no conjunto de treino empregado pelo modelo BERT, e que possivelmente explica a superioridade do modelo BERT na classe *Controle*. Em outras palavras, a habilidade de detecção do sistema ChatGPT é significativa, mas o conceito de ‘população média’ ainda é algo que não foi adequadamente modelado pela presente engenharia de *prompts* dado que o modelo ChatGPT *zero-shot* não conta com exemplos do que seria essa população.

6. Conclusões

Este artigo apresentou um primeiro estudo sobre o possível uso da ferramenta ChatGPT em uma tarefa de PLN de natureza notadamente sensível e complexa - a predição de transtorno depressivo em redes sociais - e sua comparação com um método tradicional baseado em BERT. Nossos resultados indicam que, embora ambos modelos tenham obtido acurácia média semelhante, o modelo baseado em ChatGPT pode ser considerado superior no sentido de não fazer uso de dados rotulados manualmente, enquanto o modelo BERT supervisionado exige *cópus* de treinamento anotado.

Voltando à questão do título deste artigo - BERT supervisionado ou ChatGPT *zero-shot* - propomos uma resposta indireta. Apesar dos resultados médios similares, o método *zero-shot* é melhor na classe positiva (i.e., na detecção de usuários diagnosticados), enquanto o método supervisionado é melhor na classe negativa (ou grupo de *Controle*). Como essa diferença decorre da forma como a presente tarefa computacional é definida (ou seja, como uma tarefa de distinção entre indivíduos diagnosticados e indivíduos aleatórios que representam uma população média), não é possível verificar essa questão com base no *cópus* empregado no presente estudo. É possível entretanto que a vantagem do modelo ChatGPT seja ainda mais expressiva em um cenário de classificação

dito tradicional, como o da distinção entre indivíduos depressivos e não depressivos. Um estudo desta natureza é deixado como sugestão de trabalho futuro.

Mesmo considerando-se as peculiaridades da presente definição do problema, observamos também que a fragilidade do modelo baseado em ChatGPT parece não estar tanto no modelo de língua em si, mas sim na forma como o rótulo de classe é decidido. No presente estudo, optou-se por utilizar a média simples dos escores do sistema com uso de um valor de *threshold* previamente computado para decidir se a resposta do modelo seria *Diagnosticado* ou *Controle*, mas é possível que um método mais sofisticado possa aproximar esses resultados dos obtidos pelo modelo BERT supervisionado.

Finalmente, cabe observar que o uso de métodos baseados em ChatGPT e afins pode ser menos adequado a tarefas de caracterização autoral como no presente caso, e mais adequado a tarefas de interpretação de língua natural para extração de significado textual, como análise de sentimentos ou detecção de posicionamentos [Pavan et al. 2020, Pavan and Paraboni 2022] aos moldes apresentados em [Zhang et al. 2023]. Uma iniciativa de investigação desta natureza também é deixada como sugestão de trabalho futuro.

7. Agradecimentos

Esse trabalho conta com apoio FAPESP # 2021/08213-0. Os autores agradecem ao Centro de Inteligência Artificial (C4AI-USP) e ao apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP # 2019/07665-4) e da IBM Corporation. O primeiro autor recebe apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 (# 88887.475847/2020-00).

Referências

- Ansari, L. and Ji, S. (2022). Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on computational Social Systems*.
- Aragón, M. E., López-Monroy, A. P., González-Gurrola, L. C., and y Gómez, M. M. (2019). Detecting depression in social media using fine-grained emotions. In *NAACL-2019 Proceedings*, pages 1481–1486, Minneapolis, USA. Assoc for Comp Ling.
- BigScience Workshop (2022). BLOOM (revision 4ab0472).
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.
- Burdisso, S. G., Errecalde, M., and y Gómez, M. M. (2020). t-SS3: a text classifier with dynamic n-grams for early risk detection over text streams. *Pattern Recognition Letters*, 138:130–137.
- Cacheda, F., Fernandez, D., Novoa, F. J., and Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *J Med Internet Res*, 21(6):e12554.

- Chancellor, S. and Choudhury, M. D. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *npj Digit. Med.*, 3(43).
- Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and v Goharian (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *COLING-2018*, pages 1485–1497, Santa Fe, USA.
- Coppersmith, G., Dredze, M., Harman, C., Kristy, H., and Mitchell, M. (2015). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 31–39, Denver, USA.
- da Costa, P. B., Pavan, M. C., dos Santos, W. R., da Silva, S. C., and Paraboni, I. (2023). BERTabaporu: assessing a genre-specific language model for Portuguese NLP. In *Recent Advances in Natural Language Processing (RANLP-2023)*, Varna, Bulgaria.
- da Silva, S. C., Ferreira, T. C., Ramos, R. M. S., and Paraboni, I. (2020). Data driven and psycholinguistics motivated approaches to hate speech detection. *Computación y Sistemas*, 24(3):1179–1188.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 Proceedings*, pages 4171–4186, Minneapolis, USA.
- dos Santos, W. R., de Oliveira, R. L., and Paraboni, I. (2023). SetembroBR: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*.
- dos Santos, W. R., Funabashi, A. M. M., and Paraboni, I. (2020a). Searching Brazilian Twitter for signs of mental health issues. In *12th International Conference on Language Resources and Evaluation (LREC-2020)*, pages 6113–6119, Marseille, France.
- dos Santos, W. R., Ramos, R. M. S., and Paraboni, I. (2020b). Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, 25(4):268–287.
- Flores, A. M., Pavan, M. C., and Paraboni, I. (2022). User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems*, 58(1):67–89.
- Kumar, A., Sharma, A., and Arora, A. (2019). Anxious depression prediction in real-time social data. In *Intl. Conf. on Advances in Engineering Science Management & Technology*, Dehradun, India.
- Kuzman, T., Mozetič, I., and Ljubešić, N. (2023). ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. *arXiv preprint arXiv:2303.03953*.
- Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., and Leung, H. (2020). *SenseMood: Depression Detection on Social Media*, pages 407–411. Association for Computing Machinery, New York, USA.

- Losada, D. E. and Crestani, F. (2016). A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–39, Cham. Springer.
- Losada, D. E., Crestani, F., and Parapar, J. (2017). eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In *LNCS 10456*, pages 346–360, Cham. Springer.
- Losada, D. E., Crestani, F., and Parapar, J. (2019). Overview of eRisk 2019 Early Risk Prediction on the Internet. In *LNCS 11696*.
- Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., and Schwartz, H. A. (2018). CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, USA.
- Mikolov, T., Wen-tau, S., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proc. of NAACL-HLT-2013*, pages 746–751, Atlanta, USA. Assoc for Comp Ling.
- Parapar, J., Martin-Rodilla, P., Losada, D. E., and Crestani, F. (2022). Overview of eRisk 2022: Early Risk Prediction on the Internet. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, pages 821–850, Bologna, Italy.
- Pavan, M. C., dos Santos, V. G., Lan, A. G. J., ao Trevisan Martins, J., dos Santos, W. R., Deutsch, C., da Costa, P. B., Hsieh, F. C., and Paraboni, I. (2023). Morality classification in natural language text. *IEEE transactions on Affective Computing*, 14(1):857–863.
- Pavan, M. C., dos Santos, W. R., and Paraboni, I. (2020). Twitter Moral Stance Classification using Long Short-Term Memory Networks. In *9th Brazilian Conference on Intelligent Systems (BRACIS). LNAI 12319*, pages 636–647. Springer.
- Pavan, M. C. and Paraboni, I. (2022). Cross-target stance classification as domain adaptation. In Pichardo Lagunas, O., Martínez-Miranda, J., and Martínez Seis, B., editors, *Advances in Computational Intelligence - MICAI 2022 - Lecture Notes in Artificial Intelligence vol 13612*, pages 15–25, Cham. Springer Nature Switzerland.
- Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.
- Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *arXiv preprint arXiv:2302.06476*.
- Souza, V., Nobre, J., and Becker, K. (2020). Characterization of anxiety, depression, and their comorbidity from texts of social networks. In *SBBD-2020*, pages 121–132, Porto Alegre, Brazil. SBC.
- Souza, V., Nobre, J., and Becker, K. (2021). A deep learning ensemble to classify anxiety, depression, and their comorbidity from texts of social networks. *Journal of Information and Data Management*, 12(3):306–325.

- Su, C., Xu, Z., Pathak, J., and Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(116).
- Trotzek, M., Koitka, S., and Friedrich, C. M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*.
- Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., Thirunarayan, K., Meddar, J. M., Myers, A., Pathak, J., and Hitzler, P. (2020). Multimodal mental health analysis in social media. *PLOS ONE*, 15(4):1–27.
- Zhang, B., Ding, D., and Jing, L. (2023). How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Semantic Textual Similarity: In Defense of Wordnet-Based Methods

Eduardo Corrêa Gonçalves¹

¹Escola Nacional de Ciências Estatísticas (ENCE/IBGE)

Rio de Janeiro – RJ – Brazil

eduardo.correa@ibge.gov.br

Abstract. *Wordnets have long been used as a tool for evaluating the semantic similarity between short texts. In addition to being simpler than recent deep learning approaches, methods based on wordnets offer an important advantage: they deliver results that are easy to interpret as their decisions are usually taken by considering the proximity between graph nodes. In this work, we explore a lightweight approach based on a Portuguese wordnet to solve the ASSIN 2 Semantic Textual Similarity (STS) shared task. In this task, each object of a dataset consists of a pair of Portuguese sentences annotated with its semantic score and the goal is to learn an STS model to estimate the similarity value of new, previously unseen, sentence pairs. Experiments show that our results are competitive with state-of-the-art methods in terms of mean squared error.*

1. Introduction

Semantic Textual Similarity (STS) is the task of assessing the degree of semantic equivalence between two short pieces of text [Chandrasekaran and Mago 2021, Agirre et al. 2012]. A popular application of STS can be found in Question Answering (QA) Systems [Soares and Parreiras 2020], i.e., systems that automatically answer questions posed by humans in natural language. Consider, for instance, two distinct users of a medical QA System who are interested in obtaining information about the symptoms of diabetes. The first user could ask “what are the signs of diabetes?” whereas the second might pose the question as “how can I check if I have diabetes?”. Once the questions are equivalent, the system should be capable of providing the same answer for both users.

In addition to QA Systems, there are several other important applications of STS, varying from plagiarism detection [Ferrero et al. 2017] to the comparison of product descriptions [de Lima and Gonçalves 2022]. As a result, a number of competitions (challenges and shared tasks) to promote research in STS have been run over the last few years, such as n2c2/OHNL Clinical STS Track [Wang et al. 2020], SemEval Task on STS [Agirre et al. 2012, Cer et. al 2017], and ASSIN, *Avaliação de Similaridade Semântica e Inferência Textual* (Evaluating Semantic Similarity and Textual Entailment) [Fonseca et al. 2016, Real et. al. 2019]. The second and last edition of ASSIN – which we will refer to as ASSIN 2 in the remainder of this paper – is the focus of the present work.

The organizers of ASSIN 2 shared task made available a dataset composed of about 10,000 pairs of sentences in Brazilian Portuguese (6,500 for training, 500 for validation, and 2,448 for testing). Each pair is assigned a semantic similarity score between 1.0 (the sentences are completely unrelated) and 5.0 (the sentences are

equivalent). For instance, the pair “*Um homem está tocando uma flauta*” (“A man is playing a flute”) and “*Um homem está tocando um instrumento*” (“A man is playing an instrument”) is scored with 4.5. A total of nine teams participated in the challenge whose aim was to produce the best model in terms of Pearson correlation (ρ), with Mean Squared Error (MSE) being considered as a secondary evaluation metric.

At the end of the challenge, state-of-the-art BERT-based models have shown remarkable performance, obtaining the overall best results [Fonseca and Alvarenga 2019; Rodrigues et al. 2019a; Rodrigues et al. 2019b]. However, despite their effectiveness, it is necessary to observe that BERT and its variants suffer from a drawback: they were designed to maximize predictive performance, but do not consider the comprehensibility (interpretability) of the model. This is a relevant issue for application domains in which it is necessary to determine how a model came to its conclusions, either for legal/transparency reasons or to follow ethical guidelines. For example, this is often the case of natural language processing (NLP) applications in the context of public administration, as stressed in [Anthopoulos and Wood 2021; Darrazão et al. 2023; de Lima and Gonçalves 2022].

In this paper, we explore a lightweight approach based on a Portuguese wordnet, namely Onto.PT [Gonçalo Oliveira and Gomes 2014], to solve the ASSIN 2 STS task. More specifically, the proposed strategy is based on the combination of a few traditional lexical and distributional features with semantic features computed with the utilization of Onto.PT. The choice of a wordnet-based solution was mainly motivated by the fact that this kind of structure provides the user with means for understanding how the model is generating predictions, since model’s decisions are usually taken by considering the proximity between nodes in the wordnet graph. Experiments with ASSIN 2 collection showed that our proposal achieved competitive MSE results compared with most of the state-of-the-art deep learning methods.

The remainder of the paper is divided as follows. Section 2 reviews work related to our proposal and gives a short overview on wordnets. Section 3 outlines the advantages of wordnet-based STS models and presents our proposed approach. Section 4 reports the experimental results. Concluding remarks and future directions are given in Section 5.

2. Related Work and Background

This section revises the work related to our proposal and gives an overview of wordnet concepts relevant to this paper.

2.1. Best-Performing Methods from ASSIN 2

In this subsection, we briefly describe six of the best performing algorithms developed for ASSIN 2. The model of Rodrigues et al. (2019a) ranked first place in the competition – with a Pearson ρ of 0.826 and an MSE of 0.52 – using a solution based on a pre-trained multilingual version of BERT-Base. To improve the effectiveness, they added one untrained layer of neurons, and then trained the new model using the ASSIN 2 training set along with the Brazilian Portuguese training set of the first ASSIN task (ASSIN 1).

The Stilingue team [Fonseca and Alvarenga 2019] attained the best MSE performance (0.47) and the second-best Pearson correlation score (0.817). Their proposal consists of a wide / deep learn model (based on multilingual BERT-Base and Universal Sentence Encoder-Large multilingual) combined with 18 features that describe lexical,

syntactic and semantic information from the sentences in the dataset (e.g.: jaccard similarity, negation agreement, difference in the amount of tokens between the two sentences, among others).

The method proposed in [Rodrigues et al. 2019b] consists of a stacked ensemble approach that combines the predictions generated by two models: a multilingual BERT model fine-tuned over ASSIN 1 and ASSIN 2 datasets; and a RoBERTa model fine-tuned over the automatic translation of the datasets into English. The method obtained a Pearson ρ of 0.785 (third best) and MSE of 0.59 (fourth best).

Following a different approach, more related to the present work, Santos et al. (2019) developed a traditional machine learning framework based on the evaluation of a collection of lexical, distributional, syntactic, and semantic attributes. Feature selection strategies were employed as preprocessing step to discover the subset of input features most relevant for the STS task. As a result, the original set of 71 attributes was reduced to 12, comprising only lexical and distributional features (surprisingly, none of the syntactic and semantic features were considered as relevant). The technique achieved competitive results, with Pearson ρ and MSE of 0.740 and 0.60, respectively.

Other good-performing methods were proposed by Cabezudo et al. (2019) and de Souza et al. (2019). The first fine-tuned multilingual BERT on ASSIN 2 corpus without any extra feature whilst the later trained a Siamese neural network model using various distinct features, including lexical-based, word2vec embeddings, and also incorporating similarity metrics obtained from a multilingual wordnet.

2.2. Wordnets

A wordnet is a lexical database of a given language [de Paiva et al. 2016, Fellbaum 1998]. In this kind of structure, words are organized into groups of synonymous lexical items, known as *synsets*, which are linked to each other according to their conceptual-semantic relations. Examples of such relations, among others, include “is-a”, which links more specific synsets (called *hyponyms*) to more general ones (called *hypernyms*) and “antonymy”, which indicates semantic opposition between two synsets. Princeton’s WordNet (PWN) [Fellbaum 1998], the first wordnet released, was manually created by a multidisciplinary team in the early 1990s having English as its target language. Since then, wordnets in dozens of other languages including Portuguese have been developed and successfully established [de Paiva et al. 2016].

A wordnet can be constructed as a graph where nodes are synsets and edges represent their semantic relations. Figure 1 shows an excerpt from Onto.PT [Gonalo Oliveira and Gomes 2014] – the Portuguese wordnet used in this study – showing five different synsets and the “is-a” (hyponym-hypernymy) relations between them. In this example, the topmost node represents the general concept {*cardume*, *peixe*, *peixes*} ({shoal, fish, fishes}). Three hyponym (more specific) synsets are linked to this synset, denoting three distinct kinds of fishes: {*truta*, *truite*} (trout fish), {*linguado*} (halibut), and {*bocar u*, *biquer o*, *manjuba*, *enchova*, *anchova*} (five distinct Portuguese names for anchovy fish). In the same way, {*truta-salmonada*} (salmon trout, i.e., a specific kind of trout) is linked to its hypernym {*truta*, *truite*}. Aside from relations, each synset in the wordnet hold two properties: its part of speech and a dictionary-style definition called gloss (see Figure 2).

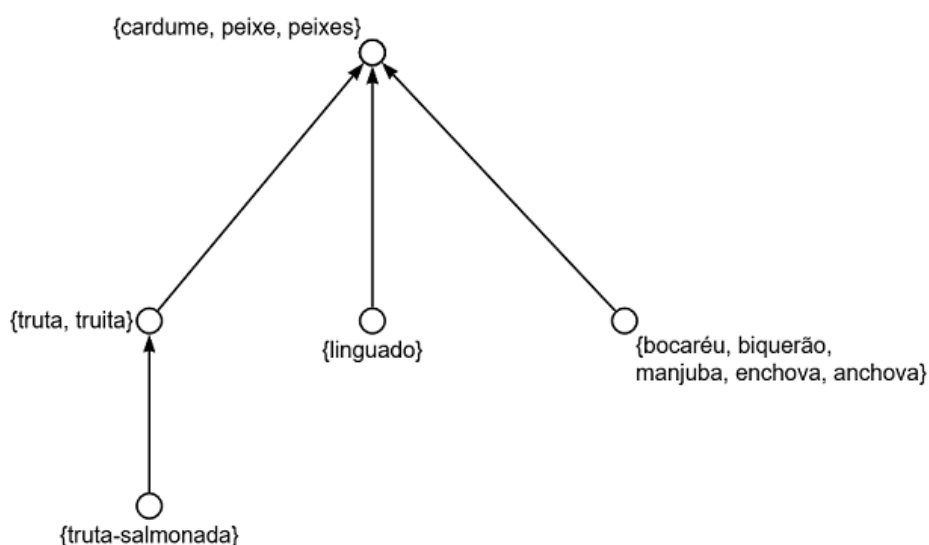


Figure 1. An excerpt from Onto.PT

{**truta-salmonada**} (substantivo): *truta com pintas rosas e carne mais avermelhada*¹.

Figure 2. A synset and its properties

Given a wordnet, there are a few different approaches for measuring the semantic similarity between two words w_1 and w_2 [Pilehvar and Navigli 2015]. Initially, it should be verified if both words belong to the same synset. If this occurs, then there is a sense of w_1 and w_2 in which they are synonymous. Otherwise, the most adopted approach is the one known as *edge counting*. According to this strategy, the similarity between w_1 and w_2 is computed by finding the shortest path between a synset containing w_1 and another synset containing w_2 in the wordnet graph. The less the number of edges in the path, the more semantically similar the words are.

The above approach can straightforwardly be extended to account for the STS task by computing the similarity between two sentences as the average of the similarity between the closest (most similar) word pairs in both sentences. This idea has been applied with good results in [Li et al. 2006, Croft et al. 2013, de Lima and Gonçalves 2022].

3. In Defense of Wordnet-based STS

When it comes to means of incorporating semantic knowledge into NLP algorithms, there are two main approaches: wordnets and word embeddings [Gonçalo Oliveira 2018]. Recently, BERT contextual embeddings [Devlin et al. 2019] and its variants have been preferred over wordnets due to their remarkable performance in a number of distinct NLP tasks – including the STS task [Fonseca and Alvarenga 2019; Rodrigues et al. 2019a; Rodrigues et al. 2019b, Wang et al. 2020].

¹ {**salmon trout**} (noun): species of trout that has pink spots and a redder meat

Although wordnets’ disadvantages compared to BERT and other embedding technologies are widely acknowledged, their advantages are rarely mentioned. Nonetheless, wordnets have appealing properties for STS applications. First, as shown in Section 2, wordnets are theoretically simple and intuitive. Aside from this, wordnets like PWN are more formalized since they have been created and maintained by experts who are responsible for grouping synsets and defining relations amongst them [Gonalo Oliveira et al. 2021].

Second and more importantly, methods based on wordnets can deliver results that are easy to interpret, as their decisions are usually taken by considering the graph topology (i.e., semantically similar concepts are located in nodes that are close to each other). On the other hand, BERT architecture was designed to maximize predictive performance but do not consider the comprehensibility (interpretability) of the model. This is a disadvantage because in certain domains, like public administration [Anthopoulos and Wood 2021; Darrazo et al. 2023; de Lima and Gonalves 2022], the ability of users to understand relevant aspects of the modeling process is also important or even required due to legal reasons, transparency issues or to follow ethical guidelines [Freitas 2014].

Motivated by these issues, in this paper we explore a lightweight wordnet-based approach to solve the ASSIN 2 STS task. Our approach is based on a small set of five features, where each takes account of either lexical / distributional similarity or semantic similarity between Portuguese sentences. The semantic features are computed with the utilization of Onto.PT [Gonalo Oliveira and Gomes 2014], the largest Portuguese wordnet. In the next subsection we describe our approach. First, we present the preprocessing steps that were performed before generating the features. Next, the features themselves are described.

3.1. The Proposed Approach

3.1.1 Preprocessing

Two preprocessing steps were carried out in the ASSIN 2 datasets: stop word removal and stemming. In the stop word removal step, pronouns, articles, prepositions, conjunctions, and linking verbs were removed from the sentences. However, adverbs were kept as they can modify the meaning of an entire sentence. The process was done with the use of the NLTK standard Portuguese stoplist [Bird et al. 2009].

Next, we assessed the coverage of Onto.PT with respect to the set of tokens (individual words or unigrams) present in the ASSIN 2 training set. We identified that from a total of 2,253 distinct tokens in the training set, only 63.43% (1,429) could be found as lexical items in Onto.PT. To mitigate this problem, we decided to submit both, the ASSIN 2 collection and Onto.PT to a stemming process employing the RSLP algorithm [Orengo and Huyck 2001]. In the stemming process, suffixes common in the Portuguese language (due to plurals or tenses) are trimmed to reduce any word to its stem. As a result, the number of distinct tokens in the ASSIN 2 training dataset was reduced from 2,253 words to 1,466 stems, where 1,364 (93.04%) could be found in Onto.PT. Table 1 summarizes the results of the stemming process. From now on, the terms stem and token will be used interchangeably in this paper, since in our approach tokens are represented by their stems.

Table 1. Summary of the stemming process

ASSIN 2 training dataset	Number of distinct tokens	Number and percentage of tokens found in Onto.PT
Before stemming	2,253	1,429 (63.43%)
After stemming	1,466	1,364 (93.04%)

3.1.2 Features

In this subsection, we present the set of five features used to build our STS model. In the definitions throughout the text, we adopted the following notation:

- t and h : the two sentences whose similarity score is to be computed with their words transformed into stems.
- Tok_t and Tok_h : the set of tokens obtained from t and h , respectively. As aforementioned, tokens are represented by the stems of the words in t and h .

Semantic Features

Two semantic features were used in our model. Both were computed according to the wordnet-based semantic similarity function presented in Equation (1).

$$F_{wordnet}(t, h) = \frac{|ExtTok_t \cap Tok_s|}{\max(|Tok_t|, |Tok_s|)} \quad (1)$$

In Equation (1), $ExtTok_t$ corresponds to Tok_t augmented with additional tokens that are somehow related to each token in Tok_t according to a wordnet. Greater values indicate higher similarity between the sentences. We employed two different approaches to determine $ExtTok_t$, which consequently led to the generation of two different semantic features, named $F_{wordnet_synonyms}$ and $F_{wordnet_hypernyms}$:

- $F_{wordnet_synonyms}$: to compute this feature, $ExtTok_t$ was generated by augmenting Tok_t with the synonyms found in Onto.PT for each of its tokens.
- $F_{wordnet_hypernyms}$: in this case, $ExtTok_t$ was generated by augmenting Tok_t with the hypernyms found in Onto.PT for each of its tokens.

Lexical and Distributional Features

Following the approach of some of the teams that participated in ASSIN 2 [Fonseca and Alvarenga 2019, Santos et al. 2019, de Souza et al. 2019], we combined the above semantic features with three additional features that explore either lexical or distributional information contained in the sentences. They are described below:

- $F_{TokensRatio}$: corresponds to the ratio of the amount of tokens (stems) in t to the amount of tokens in h . The rationale is that semantically similar sentences are expected not to have a large difference in their corresponding number of tokens.
- $F_{n-grams}$: corresponds to the cosine between the character n -grams vectors of t and h . In this work, we chose $n=5$ since our preliminary experiments found character 5-grams to be more effective than 2-grams, 3-grams, and 4-grams.
- F_{tf-idf} : the cosine of the TF-IDF vectors of the sentences. This distributional feature reflects the importance of each word stem in a sentence.

4. Results

We performed two distinct experiments. The first was carried out on the validation dataset in order to compare the performance of the features when used alone and when combined. I.e.: we created six different regression models, the first five trained with a single feature and the last one trained with the complete set of features ($F_{wordnet_synonyms}$, $F_{wordnet_hypernyms}$, $F_{TokensRatio}$, $F_{n-grams}$, and F_{tf-idf}). The models were trained using the multi-layer perceptron regressor implementation available at scikit-learn [Pedregosa et al. 2011], with default parameters, except for the maximum number of iterations (*max_iter*), which was set to 1,000. Results are shown in Table 2. It is possible to observe that, amongst the models trained with a single feature, those that performed better in the validation dataset were the model trained with F_{tf-idf} (Pearson ρ of 0.694 and MSE of 0.51) and the one trained with $F_{wordnet_synonyms}$ (Pearson ρ and MSE of 0.688 and 0.51, respectively). Nonetheless, the regression model trained with the complete set of features achieved superior performance in the validation dataset, with Pearson correlation of 0.730 and MSE of 0.46. It is worth mentioning that we also evaluated combinations of two, three, and four features, but they did not perform as effectively as the model trained with the full set of features.

Next, we conducted a second experiment on the test collection. We compared the performance of our proposed wordnet-based approach against the Pearson ρ and MSE results originally obtained by the nine teams that participated in ASSIN 2 (these results had been previously published in [Real et al. 2019]). The comparison includes the six methods presented in Section 2: ASAPPy [Santos et al. 2019], Deep Learning Brasil [Rodrigues et al. 2019b], IPR [Rodrigues et al. 2019a], NILC [Cabezudo et al. 2019], PUCPR [Souza et al. 2019], and Stilingue [Fonseca and Alvarenga 2019]. Results are presented in Table 3. In this table, the rank obtained by each method in each performance metric (Pearson and MSE) is presented in parenthesis.

Table 2. Preliminary results on the validation set

Feature(s) used to build the model	Pearson	MSE
Tokens Ratio ($F_{TokensRatio}$)	0.354 (5)	0.85 (5)
Character 5-gram ($F_{5-grams}$)	0.568 (4)	0.66 (4)
TF-IDF (F_{tf-idf})	0.694 (2)	0.51 (2)
Semantic Feature – Synonyms ($F_{wordnet_synonyms}$)	0.688 (3)	0.51 (2)
Semantic Feature – Hypernyms ($F_{wordnet_hypernyms}$)	0.117 (6)	0.96 (6)
Full wordnet-based model – all features combined ($F_{TokensRatio} + F_{5-gram} + F_{tf-idf} + F_{wordnet_synonyms} + F_{wordnet_hypernyms}$)	0.730 (1)	0.46 (1)

Table 3. Final results on the test dataset: our method versus ASSIN 2 participants

Method	Pearson	MSE
Wordnet-based model ($F_{TokensRatio} + F_{5-gram} + F_{tf-idf} + F_{wordnet_synonyms} + F_{wordnet_hypernyms}$)	0.735 (6)	0.52 (2)
ASAPPj	0.652 (9)	0.61 (7)
ASAPPy	0.740 (5)	0.60 (6)
Deep Learning Brasil	0.785 (3)	0.59 (5)
IPR	0.826 (1)	0.52 (2)
L2F/L2F INESC	0.778 (4)	0.52 (2)
LIACC	0.493 (10)	1.08 (10)
NILC	0.729 (7)	0.64 (8)
PUCPR	0.678 (8)	0.85 (9)
Stilingue	0.817 (2)	0.47 (1)

Table 3 shows that our method obtained competitive results with state-of-the-art BERT methods in terms of MSE (second best result) even though it was built using basic semantic features (proportion of synonyms and hypernyms) combined with a small set of lexical and distributional features. These results encourage us to continue investigating other, more sophisticated approaches based on wordnets to solve STS tasks.

In what follows, some drawbacks related to the use Onto.PT in STS problems will be discussed. These drawbacks might have been responsible for negative impacts on the effectiveness of the proposed method. First, it is important to state that differently from PWN and several other wordnets, Onto.PT was not handcrafted by experts. Instead, it was built through an automated process of extracting, clustering and connecting terms present in Portuguese dictionaries, thesauri, and wordnets. Consequently, it has limitations and errors. For instance, Gonçalves Oliveira and Gomes (2014) reports that an evaluation by two judges on a random sample of 774 distinct Onto.PT synsets showed that only 73.9 % of those were considered correct by both judges.

Aside from this, other relevant disadvantage of Onto.PT for STS is the fact that 56.82% of its synsets are directly connected to the root node of the graph (i.e., they do not have a hypernym). This characteristic of Onto.PT topology has hindered us from evaluating edge counting algorithms, which is the category of similarity algorithms most commonly adopted by wordnet-based STS systems [Li et al 2006, Croft et al. 2013, de Lima and Gonçalves 2022]. Another important disadvantage is that only 39.48% of synsets in Onto.PT have an associated gloss. Thus, it is not possible to employ gloss-based approaches often used for disambiguating short texts and thus improving the performance of PLN systems [Pilehvar and Navigli 2015].

5. Conclusions and Future Work

This work addressed the Portuguese STS task. We explored a lightweight wordnet-based approach that is suitable for use in domains where not only the effectiveness, but also the interpretability of the model is important. We evaluated our proposed approach on ASSIN 2 collection and achieved an MSE of 0.52 and Pearson Correlation of 0.735.

As future work, we first plan to evaluate other Portuguese and multilingual wordnets [Gonçalo Oliveira 2018; de Paiva et al. 2016]. We also intend to follow an approach similar to Rodrigues et al. (2019b) – one of the best-performing methods from ASSIN 2 – by evaluating the use of PWN over the automatic translation of the ASSIN 2 datasets into English. Since PWN is not prone to the same limitations as Onto.PT and is more complete than the other Portuguese wordnets, the translation will allow the evaluation of several distinct similarity measures [Pilehvar and Navigli 2015], including those based on edge counting, along with the incorporation of gloss-based techniques for disambiguation.

Other two topics for future research that seem to deserve special attention are the following: (i) taking into consideration not only individual words (unigrams) as tokens during the STS process, but compound words as well (bigrams and trigrams); (ii) evaluating the performance of different regression algorithms instead of MLP, prioritizing transparent and easily auditable techniques.

References

- Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity". In: Proc. of the 6th Intl' Wksp on Semantic Evaluation (SemEval-2012), ACL, p. 385–393.
- Anthopoulos, T. and Wood, M. (2021) "Automated coding of Standard Industrial and Occupational Classifications (SIC/SOC)", <https://statswiki.unece.org/display/ML/Machine+Learning+Group+2021>, June.
- Bird, S., Loper, E., and Klein, E. (2009). Natural language processing with python, O'Reilly Media Inc.
- Cabezudo, M. A. S., Inácio, M., Rodrigues, A. C., Casanova, E., and de Sousa, R. F. (2019). "NILC at ASSIN 2: Exploring Multilingual Approaches". In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 49–58.
- Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: Proc. of the 11th Intl' Wksp on Semantic Evaluation (SemEval-2017), ACL, p. 1–14.
- Chandrasekaran, D. and Mago, V. (2021). Evolution of semantic similarity: A survey. In *ACM Comput. Surv.*, 54(2), pages 41:1–41:37. ACM.
- Croft, D, Coupland, S., Shell, J., Brown, S. (2013) "A Fast and Efficient Semantic Short Text Measure", In: Proc. of the 13rd UK Workshop on Computational Intelligence (UKCI), IEEE, p. 221–227.
- Darrazão, E., Amorim, V., Oliveira, K., Gomes-Jr, L. (2023). "Engenharia e Avaliação de Features para Extração de Informação em Notas Fiscais". In: Anais da XVIII Escola Regional de Banco de Dados (ERBD), SBC, p. 80–89.
- Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), ACL, p. 4171–4186.
- Fellbaum, C. (1998). WordNet: an electronic lexical database, MIT Press, Cambridge.
- Ferrero, J., Besacier, L., Schwab, D., and Agnès, F. (2017). "CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity". In: Proc. of the 11th Intl' Wksp on Semantic Evaluation (SemEval-2017), ACL, p. 109–114.
- Fonseca, E., and Alvarenga, J. P. R. (2019). "Wide And Deep Transformers Applied to Semantic Relatedness and Textual Entailment". In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 68–76.

- Fonseca, E. R., Borges dos Santos, L., Criscuolo, M., and Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. In *Linguamática*, 8(2), pages 3–13. UMinho / UVigo.
- Freitas, A. A. (2014). Comprehensible classification models – A position paper. In *SIGKDD Explorations*, 15(1), pages 1–10. ACM.
- Gonçalo Oliveira, H. (2018). Distributional and knowledge-based approaches for computing Portuguese word similarity. In *Information*, 9(35), pages 1–21. MDPI.
- Gonçalo Oliveira, H. and Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. In *Language Resources and Evaluation*, 48(2), pages 373–393. Springer.
- Gonçalo Oliveira, H., Aguiar, F. S. S., and Rademaker, A. (2021). “On the Utility of Word Embeddings for Enriching OpenWordNet-PT”, In: Proc. of the 3rd Conf. on Language, Data and Knowledge (LDK 2021), OASICS, p. 21:1–21:13.
- Li, Y., McLean, D., Bandar, Z. A., O’Shea, J. D., Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. In *IEEE Transactions on Knowledge and Data Engineering*, 18(8), pages 1138–1150. IEEE.
- de Lima, L. S. G. and Gonçalves, E. C. (2022). “Similaridade Semântica de Nomes de Produtos Alimentícios Utilizando Wordnets do Português”. In: Proc. of the XV Seminar on Ontology Research in Brazil (ONTOBRAS 2022) and VI Doctoral and Masters Consortium on Ontologies (WTDO 2022), CEUR, p. 23–31.
- Orengo, V. M. and Huyck, C. (2001). “A Stemming Algorithm for the Portuguese Language”. In: Proc. of the 8th Symposium on String Processing and Information Retrieval, IEEE, p. 186–193.
- de Paiva, V., Real, L., Gonçalo Oliveira, H., Rademaker, A., Freitas, C., Simões, A. (2016) “An overview of Portuguese WordNets”, In: Proc. of the 8th Global WordNet Conference (GWC 2016), ACL, p. 74–81.
- Pedregosa et al. (2011). Scikit-learn: Machine learning in python. In *JMLR* 12, pages 2825–2830.
- Pilehvar, M. T. and Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. In *Artificial Intelligence*, 228, pages 95–128. Elsevier.
- Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2019). “Organizing the ASSIN 2 Shared Task”. In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 1–13.
- Rodrigues, R., Couto, P., and Rodrigues, I. (2019a). “IPR: The Semantic Textual Similarity and Recognizing Textual Entailment Systems”. In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 39–47.
- Rodrigues, R. C., da Silva, J. R., de Castro, P. V. Q., da Silva, N. F. F., Soares, A. S. (2019b). “Multilingual Transformer Ensembles for Portuguese Natural Language Tasks”. In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity

- and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 27–38.
- Santos, J., Alves, A. and Gonçalo Oliveira, H. (2019). “ASAPPy: a Python Framework for Portuguese STS”. In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 14–26.
- Soares, M. A. C. and Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. In *Journal of King Saud University - Computer and Information Sciences*, 32(6), pages 635–646. Elsevier.
- de Souza, J. V. A., Oliveira, L. E. S., Gumiel, Y. B., Carvalho, D. R., Moro, C. M. C. (2019). “Incorporating Multiple Feature Groups to a Siamese Neural Network for Semantic Textual Similarity Task in Portuguese Texts”. In: Proc. of the ASSIN2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symp. in Inf. and Human Language Technology (STIL), CEUR, p. 59–68.
- Wang, Y., Fu, S., Shen, F., Henry, S., Uzuner, O., and Liu, H. (2020). Overview of the 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity. In *JMIR Med Inform.*, 8(11):e23375. JMIR.

Contextual stance classification using prompt engineering

Felipe Penhorate Carvalho de Fonseca¹, Ivandré Paraboni¹,
Luciano Antonio Digiampietri¹

¹Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
03828-000 – São Paulo – SP – Brazil

`felipe.penhorate@gmail.com, {ivandre,digiampietri}@usp.br`

Abstract. *This paper introduces a prompt-based method for few-shot learning addressing, as an application example, contextual stance classification, that is, the task of determining the attitude expressed by a given statement within a conversation thread with multiple points of view towards another statement. More specifically, we envisaged a method that uses the existing conversation thread (i.e., messages that are part of the test data) to create natural language prompts for few-shot learning with minimal reliance on training samples, whose preliminary results suggest that prompt engineering may be a competitive alternative to supervised methods both in terms of accuracy and development costs for the task at hand.*

1. Introduction

The increasingly popular use of transformers [Vaswani et al. 2017] in NLP and related fields, and the availability of so-called Large Language Models (LLMs), machine learning tasks based on text data have undergone significant changes in how training and classification may be implemented. For a start, these advances allowed a fixed *pre-trained* language model to be reused across multiple tasks by means of fine-tuning, that is, by adjusting a general (and sometimes multilingual) model to a particular downstream task [Liu et al. 2023]. This approach, also known as “pre-train, fine-tune”, is perhaps best illustrated by the wide-spread popularity of models such as BERT [Devlin et al. 2019] in the NLP field.

Nowadays the release of progressively larger language models is commonplace and, accordingly, further advances in the field have followed. In particular, since the introduction of GPT-2 [Radford et al. 2019] and GPT-3 [Brown et al. 2020] models, we may speak of a second paradigm shift known as “pre-train, prompt and predict”. In this approach, instead of fine-tuning a pre-trained language model to address a specific task, the task itself is reformulated to serve as an input to the language model as text data [Liu et al. 2023]. For instance, in a “pre-train, prompt and predict” approach to, e.g., sentiment analysis of film reviews, we may craft a natural language instruction - or prompt - as in “[*X*] the movie is [*Z*]", in which *X* is the input text to be classified and *Z* is the label (e.g., ‘good’, ‘bad’, etc.) that we would like the model to predict. By following this method, known as *prompt engineering* [Liu et al. 2023], a sufficiently robust LLM should be able to fill in the prompt slots with the most likely tokens, which in turn would provide a sentiment class label *Z* without the need for any fine-tuning of model parameters and, crucially, requiring little or no labeled training data, in what is known as few- or zero-shot learning.

A particular NLP task that may potentially benefit from these methods is the stance classification task, which consists of identifying the point of view or judgment of an individual (e.g., in favor, against, neutral, etc.) upon a target object of interest [ALDayel and Magdy 2021]. For instance, given the target ‘vaccination’, a statement as in ‘*I believe that everyone should be vaccinated compulsory*’ conveys a stance in favor of vaccination. Stance classification may in principle seem analogous to sentiment analysis [Zhang et al. 2018]) but, unlike sentiment (e.g., positive, negative, etc.) stance (e.g., for, against, etc.) is anchored on a specific target, and stance and sentiment do not actually correlate [ALDayel and Magdy 2021].

Central to our present work, we notice that standard (i.e., supervised) approaches to stance classification will require a usually large amount of training data for every target under consideration. Thus, for instance, we may need a labeled corpus of stances towards, e.g., vaccination, and in a second project we may need a new corpus labeled with stances towards climate change, and so forth. This unlimited dependency on labeled data arguably makes stance classification an ideal testbed for prompt engineering methods.

Based on these observations, in this work we introduce a prompt-based method for few-shot text learning using stance classification as an application example. More specifically, we focus on the issue of *contextual stance classification*, that is, the task of determining the stance expressed by a given statement within a conversation thread with multiple points of view [Derczynski et al. 2017]. Our method consists of using the existing messages (which are part of the test data) as prompts for few-shot learning with minimal reliance on training samples and, in doing so, we would like to show that prompt engineering is a competitive alternative to supervised methods both in terms of accuracy and development costs.

The rest of this article is organized as follows. Section 2 discusses the concept of prompt engineering and how it compares to standard supervised learning. Section 3 describes related work in the field of stance classification. Section 4 presents the materials and methods used in our own work. Section 5 presents the results of our experiments. Finally, Section 6 draws a number of conclusions from the present work.

2. Background

In this section we introduce the issue of prompt-based learning and discuss how this compares to standard supervised approach.

Let x be an input (text), and let y be an output (label or text, for instance) based on a model $P(Y|x, \theta)$. Learning the parameters θ require a labeled set of input – > output pairs, from which we may train a model to predict the described conditional probability [Liu et al. 2023]. In a supervised approach, learning is dependent on a set of training samples to obtain the probability $P(Y|x, \theta)$. However, it is often the case that a sufficiently large train dataset is not available for the required class or domain. Moreover, as discussed in the previous sections, we notice that standard target-based stance classification will require a specific training dataset for every target topic under consideration [Mohammad et al. 2016].

As an alternative to supervised learning, the recent availability of large language models (LLMs) has enabled the use of prompt-based methods in text classification tasks.

Prompt-based learning circumvents the lack of training data by creating language models that output the probability $P(x, \theta)$ based on x by itself, that is, models that can describe an output without the need to specify what the expected output would be in a training fashion [Liu et al. 2023].

Prompt-based learning comprises of three main steps: prompt addition, answer search, and answer mapping. In what follows we briefly review each of these steps in turn. Further details are provided in [Liu et al. 2023].

Prompt addition makes use of a prompting function $f_{prompt}(x)$ to modify an input text x into a prompt $x' = f_{prompt}(x)$. This consists of applying a pre-defined template to the input x based on an input slot [X] and output slot [Z], and then filling in the slot [X] with the input text x [Liu et al. 2023], thus creating a natural language instruction (or prompt) to be submitted to the language model. There are at least two methods for modeling the input [X] and the slot [Z] as text. The first method, called *cloze prompt*, takes place when slot Z appears in the middle of the text. The second method, called *prefix prompt*, takes place when the input appears entirely before [Z].

Answer search computes the highest-score text z' that maximizes the score of the language model [Liu et al. 2023]. This involves defining a set of permissible values Z for z , and then using a function $f_{fill}(x_0, z)$ to fill in the location [Z] in prompt x_0 with the candidate answer z . In classification tasks, we may define, e.g., $Z = \{excellent, good, OK, bad, horrible\}$ to represent a set of possible classes $Y = \{++, +, -, --\}$. A prompt is said to be an *answer prompt* when it fills in the output slot correctly. [Liu et al. 2023].

Finally, answer mapping establishes a mapping between the computed answer z' and the target output value y . This step may in some cases be trivial but, since multiple answers may result in the same output (e.g., ‘bad’, ‘very bad’, ‘horrible’ etc. may all be mapped onto a ‘0’ class label in a particular application), it is often necessary to establish a mapping from z' to y [Liu et al. 2023].

3. Related Work

Stance classification has been established as a major research topic in the NLP field since the SemEval stance detection shared task series [Mohammad et al. 2016, Derczynski et al. 2017] in 2016-2017, followed by RumourEval 2019 [Gorrell et al. 2018]. In what follows we briefly review these initiatives and the best-performing participant systems in each task.

SemEval 2016 Task 6 [Mohammad et al. 2016] introduced two stance detection tasks by providing a stance corpus of tweets in the English language. Task A addressed stance classification in a standard supervised setting, and Task B addressed the task in an unsupervised fashion. The SemEval 2016 corpus consisted of 4,163 tweets conveying a stance (for, against, or neutral) towards five target topics (Atheism, Climate Change, Feminist Movement, Hillary Clinton, and Abortion Legalization). An additional, unlabeled topic (Donald Trump) was used in (unsupervised) task B.

The three best-performing participant systems in SemEval 2016 Task 6 [Mohammad et al. 2016] were Mitra [Zarrella and Marsh 2016], Pkudblab [Wei et al. 2016] and Takelab [Tutek et al. 2016]. Mitra’s approach was based on a

recurrent neural network, whereas Pkudblab [Wei et al. 2016] used a convolutional neural network. Takelab, on the other hand, took a different approach by using an ensemble of models created with the aid of a genetic algorithm.

Of particular interest to the present work, SemEval 2017 Task 8 [Derczynski et al. 2017], also known as RumourEval, introduced a novel Twitter dataset for stance classification that included contextual information represented by rumors associated with the stance target, and which could be used as an aid to the classification task. The RumourEval corpus is divided into pre-defined training and a test subsets. The training portion contains 297 conversations about 8 rumors discussed across 297 *tweets* that initiated a conversation thread, and 4,222 answers, making 4,519 *tweets* in total. The test dataset has 28 conversations, being 20 about the same rumors introduced in the training dataset, and 8 are about different rumors. Table 1 presents the class distribution of the train and test datasets.

Table 1. RumourEval 2017 class label distribution.

	Support	Deny	Query	Comment
Training	910	344	358	2,907
Test	94	71	106	778

The three best-performing participant systems in SemEval 2017 Task 8 were Turing [Kochkina et al. 2017], UWaterloo [Bahuleyan and Vechtomova 2017] and ECNU [Wang et al. 2017]. Turing proposed an approach based on a recurrent neural network with LSTM neurons and additional features derived from the training data. UWaterloo [Bahuleyan and Vechtomova 2017] based their approach mainly on feature selection and engineering, some of which manually curated with external knowledge provided by annotators, and using a XGBoost classifier. ECNU [Wang et al. 2017] combined an *ensemble* approach with hierarchical training to take advantage of the contextual information provided.

SemEval 2019 Task 7 [Gorrell et al. 2018] (also known as RumourEval 2019) improved upon the original RumourEval task definition by adjusting a number of issues found in the original dataset, and by adding data from Reddit. The three best-performing participant systems were BLCU NLP [Yang et al. 2019], BUT-FIT [Fajcik et al. 2019], and eventAI [Li et al. 2019]. BLCU NLP fine-tuned a Generative Pre-Trained Transformer (GPT) for contextual stance classification taking as an input the entire conversation history, and not only the target tweet. To this end, the conversation history was submitted to the model as a natural language prompt with tweets divided by separators, and the model was subsequently fine-tuned using a fully connected layer that followed the GPT layers. BUT-FIT’s used a fine-tuned BERT model prompted with a contextual representation comprising the text that generated the conversation thread (i.e., the first text in a conversation), and the texts that appeared before and after it in the conversation. The eventAI approach, by contrast, did not use any LLM, proposing instead a recurrent neural network approach based on LSTM neurons alongside a rule-based model.

After the initial SemEval and RumourEval shared tasks, multiple stance classification datasets and models have been publicly released. These include, for instance, studies devoted to Arabic [Alhindi et al. 2021, Jaziriyani et al. 2021], Portuguese

[Won and Fernandes 2022], German [Gohring et al. 2021], and multilingual scenarios [Chen et al. 2022]. Moreover, although most studies are purely text-based, the issue of multimodal stance classification (e.g., combining text and social media relations or other knowledge sources) has also been investigated [Sakketou et al. 2022]. We notice also that some of these resources are considerably larger than the original SemEval corpus. This is the case, for instance, of the P-Stance corpus in [Li et al. 2021], comprising over 21k labeled tweets.

Finally, we notice that all of the above studies, including those that used an LLM in their architecture, addressed the issue of contextual stance classification in a standard supervised fashion, that is, none of them addressed the task using zero- or few-shot prompt engineering. Examples of this kind are only beginning to emerge in the field, and include, for example, [Yin et al. 2019, Zhang et al. 2023].

4. Materials and Methods

We envisaged an experiment in prompt-based learning to address the task of contextual stance classification as described in the previous sections. In what follows, we outline the materials and methods employed in the present work.

Our experiment makes use of the contextual stance data provided by the SemEval 2017 Task 8 corpus [Derczynski et al. 2017]. The corpus consists of a series of conversation threads in which individual messages may either *Support*, *Deny*, *Query* or *Comment* the root statement. This structure is illustrated in Figure 1, in which each example u_i consists of an input text x_i , an output y_i and a context C_i , where C_i is every u_j that occurs before the current example in the conversational tree.

SDQC support classification. Example 1:

u1: We understand there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News [**support**]
u2: @u1 not ISIS flags [**deny**]
u3: @u1 sorry - how do you know it's an ISIS flag? Can you actually confirm that? [**query**]
u4: @u3 no she can't cos it's actually not [**deny**]
u5: @u1 More on situation at Martin Place in Sydney, AU -LINK- [**comment**]
u6: @u1 Have you actually confirmed its an ISIS flag or are you talking shit [**query**]

SDQC support classification. Example 2:

u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shooting #StandforCanada -PICTURE- [**support**]
u2: @u1 Apparently a hoax. Best to take Tweet down. [**deny**]
u3: @u1 This photo was taken this morning, before the shooting. [**deny**]
u4: @u1 I don't believe there are soldiers guarding this area right now. [**deny**]
u5: @u4 wondered as well. I've reached out to someone who would know just to confirm that. Hopefully get response soon. [**comment**]
u4: @u5 ok, thanks. [**comment**]

Figure 1. Two ‘support’ classification instances from [Derczynski et al. 2017].

In this scenario, our present approach makes use of the existing conversation thread to implement prefix prompt addition. More specifically, given an user who authored a sample message u_i , the prompt is introduced by using a structure as follows.

*This is a conversation between some friends about an article in Twitter.
They decided that they can only support, deny, query or add a comment
about the article.*

The introductory statement is followed by the context C of the current example, its corresponding text [X], and the answer [Y] provided by the model, in the form “[C][X]. A:[Y]”. The context itself comprising a series of (few-shot) query-answer example pairs in the format “**Q:** User said [X]. **A:** User wants to [Y] the article” according to the structure of the conversation. In our approach, all available contextual messages are taken as learning prompts and, if necessary, additional prompts are created as discussed below.

An example of the complete prompt structure is illustrated in Figure 2, in which the current text is shown in green, the expected answer [Z] appears in red, and the context [C] appears in blue. The [Y] labels for each example are shown as [Y].

This is a conversation between some friends about an article in Twitter. They decided that they can only support, deny, query or add a comment about the article.

Q: John said “We understand there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News”

A: John wants to [support] the article

Q: Benjamin answered John “[mention] not ISIS flags”

A: Benjamin wants to [deny] the article

Q: Charlotte answered John “[mention] sorry - how do you know it’s an ISIS flag? Can you actually confirm that?”

A: Charlotte wants to [query] the article

Q: Amelia answered Charlotte “[mention] no she can’t cos it’s actually not”

A: Amelia wants to [deny] the article

Q: Mia answered John “[mention] More on situation at Martin Place in Sydney, AU –LINK–”

A: Mia wants to [comment] the article

Q: Paul answered John “[mention] Have you actually confirmed its an ISIS flag or are you talking shit”

A: Paul wants to [query] the article

Figure 2. A prompt structure example and its expected output (in red).

In this representation, whenever an input is a reply to another message, the text changes to “**Q:** UserA answered UserB [X]. **A:** UserA wants to [Y] the article”, in which User is a placeholder for the user to whom a message is assigned. To this end, the original username of each individual is replaced with a name within a pre-defined set of possible names $N = \{John, Paul, Lily, Noah, Olivia, James, Lucas, Emma, Amelia, Henry, Liam, Charlotte, Elijah, Ava, William, Sophia, Benjamin, Isabella, Mia, Evelyn, Theodore, Harper\}$.

The goal of the classifier is to complete the last query-answer pair in the sequence of the conversation with the intended class prediction, and for that reason it is imperative that context [C] includes at least one example of each possible answer. Since not all corpus

conversations are complete in this way, if necessary the context will be expanded with additional query-answer pair taken from the training portion of the RumourEval corpus, which is otherwise discarded. These additional samples are selected from messages with a time prior to the time of the current message and according to cosine similarity.

By following this procedure, prompts were engineered for every test instance in the RumourEval corpus, and then submitted to the OpenAI GPT 3.5 *text-davinci-003* model with default temperature. This choice was partially motivated by its ability to handle up to 4,097 tokens as an input, which is sufficiently large to handle most of the prompts generated by the present method.

As a generative model, GPT 3.5 may naturally provide answers in multiple formats and, accordingly, some form of answer mapping is called for. In the present work, answers are mapped onto *Deny*, *Support*, or *Query* class labels according to the presence of certain keywords, or otherwise mapped onto *Comment* class labels as summarized in Table 2.

Table 2. Keyword-based answer mapping.

Class	Keywords
Deny	deny, denies, denying
Support	back up, reinforce, support
Query	query, querying, queries
Comment	none of the above

5. Results

Table 3 presents RumourEval test data F1 results obtained by our prompt engineering approach, and by the two top-performing systems at RumourEval. The best results for each class are highlighted.

Table 3. RumourEval F1-score test results.

Approach	Overall	Comment	Deny	Query	Support
Turing [Kochkina et al. 2017]	0.43	0.87	0.00	0.46	0.40
UWaterloo [Bahuleyan and Vechtomova 2017]	0.45	0.87	0.06	0.49	0.40
Our work	0.47	0.76	0.37	0.54	0.22

Results in Table 3 show that our current work, although only outperforming the baseline systems in two individual tasks (*Deny* and *Query*), obtained overall highest F1-scores among the systems under evaluation. Moreover, our work was the only system capable of handling - albeit still in a limited fashion - the more challenging *Deny* task. This outcome, and the observation that our work, unlike the two baseline systems, does not require training data, suggest that the use of prompt-based methods for contextual stance classification may represent a compelling alternative to standard approaches that rely on model supervision.

6. Conclusion

This article introduced a few-shot method to contextual stance classification using test messages available from the current conversation thread (i.e., within which the target message occurs) to prompt a large generative model, with results that show improvement over the two best-performing participant systems at RumourEval. In addition to that, results also show a considerably higher accuracy in handling so-called ‘Deny’ statements if compared to previous work, which is a likely benefit of using a large language model for the task.

More importantly, unlike previous work in the field, we notice that the current results were obtained in a few-shot fashion, that is, with no reliance on a large training dataset. Thus, if taking into account its underlying development costs, the present approach affords a significant advantage over existing methods. This is particularly the case of manual corpus annotation, a task that, in standard stance classification, would normally have to be performed for every single target topic of interest, with substantial costs that are presently negligible.

The present work leaves a number of opportunities open to investigation. First, we notice that the current model may be further assessed using the extended RumourEval 2019 dataset in [Gorrell et al. 2018], or other similar resources. Second, we may consider alternative prompt engineering methods including, for instance, enriching the prompt instructions with external knowledge about the conversation topic (e.g., from news articles, Wikipedia, etc.) Moreover, the present approach may in principle be applied to other text classification tasks based on contextual information including, for instance, sarcasm or sentiment detection.

References

- ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.
- Alhindi, T., Alabdulkarim, A., Alshehri, A., Abdul-Mageed, M., and Nakov, P. (2021). AraStance: A multi-country and multi-domain dataset of Arabic stance detection for fact checking. In *4th Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 57–65, Online. Assoc. for Computational Linguistics.
- Bahuleyan, H. and Vechtomova, O. (2017). UWaterloo at SemEval-2017 task 8: Detecting stance towards rumours with topic independent features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 461–464, Vancouver, Canada. Association for Computational Linguistics.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.
- Chen, N., Chen, X., and Pang, J. (2022). A multilingual dataset of covid-19 vaccination attitudes on twitter. *Data in Brief*, 44:108503.

- Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., and Zubiaga, A. (2017). SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.
- Devlin, J., Chang, M., 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Fajcik, M., Smrz, P., and Burget, L. (2019). BUT-FIT at SemEval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Gohring, A., Klenner, M., and Conrad, S. (2021). DeInStance: Creating and evaluating a german corpus for fine-grained inferred stance detection. In *17th Conference on Natural Language Processing (KONVENS 2021)*, pages 213–217, Düsseldorf, Germany. KONVENS 2021 Organizers.
- Gorrell, G., Bontcheva, K., Derczynski, L., Kochkina, E., Liakata, M., and Zubiaga, A. (2018). Rumoureal 2019: Determining rumour veracity and support for rumours.
- Jaziriyani, M. M., Akbari, A., and Karbasi, H. (2021). ExaASC: A General Target-Based Stance Detection Corpus in Arabic Language. In *11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 424–429, Mashhad, Iran. IEEE.
- Kochkina, E., Liakata, M., and Augenstein, I. (2017). Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.
- Li, Q., Zhang, Q., and Si, L. (2019). eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 855–859, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Li, Y., Sosea, T., Sawant, A., Nair, A. J., Inkpen, D., and Caragea, C. (2021). P-stance: A large dataset for stance detection in political domain. In *Findings of ACL-IJCNLP-2021*, pages 2355–2365, Online. Assoc. for Computational Linguistics.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).
- Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
- Sakketou, F., Lahnala, A., Vogel, L., and Flek, L. (2022). Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion. In *LREC-2022 proceedings*, pages 3798–3808, Marseille, France. ELRA.
- Tutek, M., Sekulić, I., Gombar, P., Paljak, I., Čulinović, F., Boltužić, F., Karan, M., Alagić, D., and Šnajder, J. (2016). TakeLab at SemEval-2016 task 6: Stance classification in tweets using a genetic algorithm based ensemble. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 464–468, San Diego, California. Association for Computational Linguistics.
- 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*, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Wang, F., Lan, M., and Wu, Y. (2017). ECNU at SemEval-2017 task 8: Rumour evaluation using effective features and supervised ensemble models. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 491–496, Vancouver, Canada. Association for Computational Linguistics.
- Wei, W., Zhang, X., Liu, X., Chen, W., and Wang, T. (2016). pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California. Association for Computational Linguistics.
- Won, M. and Fernandes, J. (2022). SS-PT: A stance and sentiment data set from portuguese quoted tweets. In *PROPOR-2022 proceedings*, pages 110–121, Fortaleza, Brazil. Springer.
- Yang, R., Xie, W., Liu, C., and Yu, D. (2019). BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.
- Zarrella, G. and Marsh, A. (2016). MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.
- Zhang, B., Ding, D., and Jing, L. (2023). How would Stance Detection Techniques Evolve after the Launch of ChatGPT?
- Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.

Classificação de gêneros a partir de letras de músicas em português

Matheus Bastos de Oliveira¹, João Baptista de Oliveira e Souza Filho¹

¹Programa de Engenharia Elétrica

Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brasil

{oliveiraa.maatheus, jbfilho}@poli.ufrj.br

Abstract. *Associating songs with genres is not easy. The subjectivity and diversity in musical works make assigning unequivocal labels a challenging task. However, textual features can aid in characterizing genres. This work proposes a system for classifying song lyrics in Portuguese, exploring Deep Learning models, such as LSTM networks and Transformers. It also includes more straightforward strategies like Logistic Regression classification of TF-IDF-generated embeddings. Experiments demonstrated that the Transformer model achieved the best performance, reaching an accuracy of 61.6% for ten music genres.*

Resumo. *Associar canções a gêneros não é fácil. A subjetividade e diversidade das obras musicais tornam a atribuição de rótulos inequívocos uma tarefa desafiadora. Porém, atributos textuais podem contribuir para a caracterização de gêneros. Este trabalho propõe um sistema para classificar letras de músicas em português, explorando modelos de Aprendizado Profundo, tais como Redes LSTM e Transformers. São também incluídas estratégias mais simples como a classificação por Regressão Logística de representações geradas por TF-IDF. Experimentos demonstraram que o modelo Transformer apresentou o melhor desempenho, alcançando uma acurácia de 61,6% para dez gêneros musicais.*

1. Introdução

A música, enquanto prática artística e cultural, tem como essência a diversidade e a pluralidade. Neste sentido, muitos fatores podem influenciar sua concepção, como, por exemplo, o seu local de origem, seu enquadramento cronológico e o seu contexto etnográfico.

A fim de atender tal diversidade, as obras musicais são comumente categorizadas em gêneros, de acordo com as similaridades observadas entre elas. Entretanto, tal tarefa pode ser desafiadora e demandante, uma vez que os estilos musicais podem inspirar e gerar novas formas de música com características específicas. Entre alguns exemplos têm-se: o “pop-rock”, que une características dos gêneros “pop” e “rock”; a “bossa nova”, a qual pode ser considerada tanto um gênero musical próprio quanto um subgênero do “samba” ou do “jazz”; e a “pisadinha”, que é um gênero recente, surgido em meados dos anos 2000, oriundo do “forró” [Vicente 2022]. Um outro ponto problemático é o fato de uma canção ser uma expressão artística complexa, podendo admitir diferentes interpretações, o que dificulta a sua caracterização plena por rótulos pré-determinados.

Uma vez que uma canção é uma conjugação de dois tipos de linguagens: a verbal e a musical, aspectos específicos podem estar presentes tanto em percepções auditivas, como melodia, ritmo e harmonia, como em elementos textuais [Bonds 2018].

Neste contexto, as técnicas de Processamento de Linguagem Natural (NLP) têm tornado mais factível a automatização da tarefa de identificação de gêneros a partir das letras de músicas. Tal processo pode ser útil para diferentes aplicações, tais como a indexação, recomendação e a distribuição de músicas, cada vez mais relevantes à vida digital.

Desenvolver um modelo de NLP dedicado a esta tarefa presume a coleta de um volume expressivo de letras. Em particular, o português, apesar de ser significativamente falado ao redor do mundo, carece de maiores *corpora* e de modelos específicos. Motivado por tais questões, este trabalho busca apresentar um conjunto de dados balanceado, especialmente elaborado para a classificação de gêneros musicais através de letras de canções em português, e disponibilizado em um repositório público [de Oliveira 2023]. Adicionalmente, é proposto um modelo de classificação automática que opera com gêneros tipicamente brasileiros e pouco observados na literatura, os quais são identificados apenas por meio do conteúdo semântico contido nas letras, portanto de forma independente ao ritmo segundo o qual a música é executada.

Neste estudo, considerando-se diferentes métodos de geração de representações distribuídas das palavras na forma de vetores numéricos, processo conhecido como “geração de *embeddings*”, foram avaliados algoritmos promissores para o tratamento do problema, tais como as Redes Neurais do tipo *Long Short-Term Memory* (LSTM) [Hochreiter and Schmidhuber 1997] e os *Transformers* [Vaswani et al. 2017], bem como a classificação por Regressão Logística [Hastie et al. 2009], uma alternativa mais simples, a fim de melhor relacionar a complexidade dos modelos adotados e o desempenho obtido.

Este artigo é estruturado da seguinte maneira: a Seção 2 realiza uma breve revisão bibliográfica sobre o tema de classificação de gêneros musicais; a Seção 3 descreve a base de dados e o seu pré-processamento, bem como aponta algumas características estatísticas de interesse dos dados; a Seção 4 apresenta a metodologia utilizada para o projeto dos modelos, descrevendo as arquiteturas e os processos de geração de *embeddings* empregados. Por fim, a Seção 5 exhibe os resultados obtidos, enquanto a Seção 6 discute as conclusões.

2. Trabalhos Relacionados

Trabalhos sobre a classificação automática de músicas a partir de sinais de áudio são frequentes na literatura. Em [Jeong and Lee 2016] foi proposta uma estrutura para o aprendizado de características temporais discriminantes que alcançou 63% de acurácia para a identificação de 10 gêneros, quando combinada com o aprendizado convencional de características espectrais. De forma similar, uma acurácia de 85% é reportada por [da Silva Muniz and de Oliveira e Souza Filho 2023], o qual considerou a geração de 81 atributos específicos. Cabe destacar que um estudo subjetivo descrito em [Gjerdingen and Perrott 2008] sinaliza 70% de acerto para humanos quando ouvidas amostras com uma duração de 3 segundos.

Quando consideradas letras de canções em português, em [Guimarães et al. 2020] foram realizados experimentos considerando 6000 canções e 7 gêneros: forró, gospel, MPB, samba, sertanejo, bossa nova e axé. Cada canção foi transformada num vetor de inteiros em que cada componente é o índice no vocabulário de suas 200 primeiras palavras. Em seguida, foram avaliados seis modelos de classificação: LSTM, FastText, *eXtreme Gradient Boosting*, *Random Forest* (RF), Árvore de Decisão e *Multilayer Perceptron*. O melhor resultado foi obtido pelo modelo LSTM (50% de acurácia), ao considerar uma ca-

mada de *Embedding* de 100 dimensões e treinamento explorando as técnicas de *Dropout* e *Gradient Clipping*. Em [de Araújo Lima et al. 2020], foi apresentado um conjunto de dados com cerca de 138 mil canções brasileiras e 14 gêneros. Os experimentos com ele realizados exploraram os modelos *Support Vector Machine* (SVM), RF e LSTM Bidirecional (BiLSTM), cada um associado a diferentes técnicas para a geração dos *embeddings*. Os melhores resultados foram obtidos a partir da combinação da rede BiLSTM com o modelo *Wang2Vec* pré-treinado em português, que alcançou 48% na média do F1-score inferido para cada classe. Em [Pimenta and Pugliesi 2022], foram considerados 3 gêneros: sertanejo, MPB e funk, e 12 mil músicas, realizando-se a vetorização das letras através da técnica TF-IDF. Para a classificação, foram avaliados os modelos *Linear Discriminant Analysis*, *k-Nearest Neighbors*, SVM, Árvore de Decisão, RF e Regressão Logística, o último de melhor eficácia, que atingiu uma acurácia de 80%. Uma tendência natural observada nestes trabalhos é que os modelos com um menor número de gêneros obtiveram um melhor desempenho, visto resolverem tarefas mais simples.

3. Base de Dados

Nesta seção são apresentadas as etapas envolvidas na geração e tratamento dos dados.

3.1. Coleta dos Dados

A base de dados [Neisse 2022] explorada neste trabalho foi coletada no repositório “Kaggle”, comunidade virtual de ciência de dados, e integra letras de 379,893 canções de 4,239 artistas, das quais cerca de 41% delas estão em português, e foram portanto consideradas para este trabalho, exceto quando dispunham de campos com dados faltantes.

Esta base se encontra dividida em duas tabelas distintas. A primeira, “*lyrics-data*”, relaciona cada letra de música a seu nome e idioma, bem como apresenta os endereços *web* para a letra e para o perfil do artista; a outra, “*artists-data*”, relaciona cada artista com os seus gêneros musicais, o número de canções que possui, sua pontuação de popularidade inferida pelo número de acessos ao site e o endereço *web* do seu perfil.

3.2. Pré-processamento e Formação da Base

A preparação da base de dados para a classificação envolveu os seguintes passos: para cada letra presente na *lyrics-data*, foi utilizado o endereço *web* para o perfil do artista como item de busca dos gêneros a ele associados na *artists-data*. Em seguida, mantiveram-se apenas as colunas correspondentes à letra (atributos) e ao primeiro - por ser o mais relevante ao artista - desses gêneros (alvo). A fim de expressar um bom compromisso quanto a variedade de gêneros e o quantitativo de canções disponível por gênero, foram enumerados e selecionados apenas os 10 gêneros com a maior quantidade de dados (forró, funk carioca, gospel, MPB, pagode, pop-rock, rock, romântico, samba e sertanejo). Para balancear a base, foi utilizada a técnica de *undersampling*. Com efeito, notou-se que dentre estes gêneros, o que possuía menos letras associadas contava com 3990 canções, assim foi assumida tal quantidade de letras para todos os demais gêneros. Logo após, foi realizada a divisão da base em dados de treino (para o ajuste dos parâmetros internos dos modelos), validação (para a sintonia de seus hiperparâmetros) e teste (para a avaliação final e comparação entre modelos), de forma aleatória, em proporções de, respectivamente, 72%, 8% e 20% do total de dados. Por fim, foram identificadas e removidas 12 letras cujos conteúdos não representavam canções reais, com conteúdos como “Essa música é

INSTRUMENTAL”, “(Mensagem)” e “Introdução”. Como tal quantitativo não é substancial, assumiu-se que a base continuou balanceada, dispondo-se de 3986 músicas de pop-rock, 3987 de MPB, 3988 de rock, 3989 de forró, funk e gospel, e 3990 de pagode, romântico, samba e sertanejo.

3.3. Análise Exploratória

A Tabela 1 descreve a quantidade de palavras, versos e estrofes médios por canção, bem como os valores mínimo, médio e máximo de palavras únicas por canção; e a Tabela 2, as 3 palavras mais comuns ao longo das letras, cada uma com a sua respectiva taxa de ocorrência (T_O), dada pela razão entre a quantidade de ocorrências de uma dada palavra ao longo de todas as canções de um gênero pelo total de canções deste gênero.

Cabe destacar que todas as análises não diferenciaram letras maiúsculas de minúsculas. Relativamente à Tabela 2, foram excluídas as *stopwords* indicadas pela biblioteca NLTK [Bird et al. 2009], bem como pontuações e as palavras “ai”, “vou”, “vai”, “vem”, “pra”, “pro”, “tô”, “tá”, “tão”, “mim”, “ti”, “então”, “lá”, por não se acreditar que contribuiriam significativamente para a discriminação dos gêneros.

Tabela 1. Comprimentos médios e palavras únicas por canção

Gênero	Palavras	Versos	Estrofes	Mínimo	Média	Máximo
Forró	142,1 ± 89,6	25,1 ± 16,0	5,7 ± 4,3	13	70,1	459
Funk	213,3 ± 131,1	35,9 ± 21,6	7,5 ± 5,1	6	86,2	660
Gospel	130,4 ± 73,6	22,0 ± 11,3	4,7 ± 3,1	11	68,4	571
MPB	126,9 ± 81,3	25,0 ± 16,2	4,5 ± 4,2	4	68,0	643
Pagode	144,6 ± 84,9	25,4 ± 14,7	4,9 ± 3,7	13	73,1	487
Pop-Rock	160,5 ± 82,8	30,2 ± 14,8	6,3 ± 4,2	4	75,3	466
Rock	153,3 ± 73,2	28,3 ± 14,1	6,1 ± 4,3	6	75,2	492
Romântico	128,1 ± 62,8	23,8 ± 11,8	4,8 ± 2,9	5	66,3	378
Samba	122,5 ± 70,0	23,6 ± 13,7	4,5 ± 4,1	11	68,2	309
Sertanejo	158,0 ± 65,8	27,2 ± 12,4	5,8 ± 3,9	15	86,9	307

É interessante observar na Tabela 1 o quanto os gêneros se diferenciam no comprimento médio de suas canções. Neste quesito, cabe destaque ao funk, que apresenta, para as três medidas realizadas, o valor mais expressivo e distinto dos demais. Por outro lado, nota-se uma grande semelhança entre os trios de valores apresentados pelos gêneros MPB, romântico e samba. Além disso, é surpreendente a variabilidade dos gêneros, que contêm alguns exemplos de canções com mais de 300 palavras únicas. Por outro lado, há canções com um número muito pequeno de palavras. Um caso digno de nota é a música do gênero MPB “Tutano” do artista Walter Franco, cuja letra é “Quem tem tutano, tutano tem. Quem não tem tutano, tutano não tem.”, formada portanto por apenas 4 palavras únicas.

Pela Tabela 2 constata-se que o gênero gospel é o que possui uma maior disparidade quanto às palavras mais frequentes em suas letras, visto “Deus”, “senhor” e “Jesus” não figurarem como resultado de nenhum outro estilo. Além disso, cabe destacar a grande similaridade entre as palavras mais frequentes dos demais estilos, demonstrando que o uso de palavra(s)-chave(s) é ineficaz para uma determinação direta do gênero da música, ressaltando assim a não-trivialidade do problema.

Tabela 2. Palavras mais frequentes por canção

Gênero	Palavra	T _O	Gênero	Palavra	T _O
Forró	amor	1,427	Funk	quero	0,788
	quero	0,684		quer	0,709
	coração	0,623		amor	0,623
Gospel	Deus	1,808	MPB	amor	0,851
	senhor	1,021		tudo	0,466
	Jesus	0,948		vida	0,455
Pagode	amor	1,475	Pop-Rock	tudo	0,872
	quero	0,701		amor	0,767
	gente	0,583		quero	0,728
Rock	tudo	0,809	Romântico	amor	1,696
	quero	0,693		quero	0,628
	amor	0,601		coração	0,607
Samba	amor	1,094	Sertanejo	amor	1,270
	samba	0,489		vida	0,595
	vida	0,431		coração	0,591

Por fim, a Tabela 3 indica a proporção percentual de palavras do vocabulário de cada gênero (disposto nas linhas) que também pertencem ao vocabulário de um outro estilo (disposto nas colunas). Os gêneros estão sinalizados conforme se segue: forró (0), funk (1), gospel (2), MPB (3), pagode (4), pop-rock (5), rock (6), romântico (7), samba (8) e sertanejo (9). Para facilitar a análise, o menor e o maior valor observados para cada gênero (linha) são destacados com as cores vermelha e azul, respectivamente.

Tabela 3. Matriz de coocorrência de vocabulários

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0)	100	50,4	39,1	54,7	46,7	51,6	54,5	45,6	52,7	54,6
(1)	36,5	100	29,7	39,4	36,8	41,6	43,1	32,8	38,0	39,2
(2)	43,4	45,6	100	51,1	41,7	50,1	51,9	45,5	48,4	50,1
(3)	41,7	41,4	35,0	100	37,8	48,4	50,8	38,6	48,2	45,2
(4)	51,3	55,9	41,3	54,6	100	54,7	56,4	46,3	55,3	53,8
(5)	42,1	46,9	36,8	51,8	40,6	100	56,2	40,0	46,1	46,3
(6)	41,3	45,2	35,5	50,6	38,9	52,3	100	38,6	44,9	45,3
(7)	55,1	54,9	49,5	61,3	50,9	59,3	61,5	100	59,2	61,0
(8)	43,9	43,8	36,3	52,7	41,8	47,1	49,3	40,8	100	48,6
(9)	47,4	47,0	39,2	51,5	42,4	49,2	51,8	43,8	50,7	100

Nessa tabela verifica-se que a coluna (2) é a que contém os menores valores para todas as linhas, exceto a própria (2), sinalizando que o gospel é o estilo de menor coincidência com os demais vocabulários. Por sua vez, a coluna (6) mostra que o rock é o gênero de maior similaridade de vocabulário com os demais, exceto forró e samba. Cabe destacar que nem sempre as relações são recíprocas, ou seja, tal matriz não é simétrica. Por exemplo, há muitas palavras do gênero rock nas letras do gênero romântico (61,5%), porém há poucas do gênero romântico nas letras de rock (38,6%).

4. Metodologia

Nesta seção é apresentado breve resumo dos modelos avaliados. O LSTM é de uma arquitetura capaz de armazenar padrões de interesse ao longo de intervalos de iteração arbitrários. Por sua vez, o *Dropout* [Srivastava et al. 2014]: é uma estratégia de regularização que consiste em excluir aleatoriamente um percentual (“taxa de *Dropout*”) das conexões de entrada e das conexões recorrentes a cada iteração de treinamento. O LSTM Bidirecional (BiLSTM) consiste em uma variante LSTM com duas camadas, com fluxos de informações em direções contrárias. Os *Bidirecional Encoder Representations from Transformers* [Devlin et al. 2019] (BERT) representam modelos pré-treinados, formados por camadas empilhadas de *Transformers*. Operam realizando o mapeamento das palavras do texto em *embeddings* que são submetidos a um cabeçote interno de classificação [HuggingFace 2018]. O modelo utilizado neste trabalho foi o BERTimbau-Base [Souza et al. 2020]. A estratégia BERT + Regressão Logística explora o modelo BERT pré-treinado e sem ajuste fino, que é utilizado apenas para a geração dos *embeddings*, submetidos a um classificador baseado em Regressão Logística. Por fim, o esquema *Term Frequency - Inverse Document Frequency* [Salton and Buckley 1988] (TF-IDF) + Regressão Logística se refere aos experimentos em que os *embeddings* são gerados através da técnica TF-IDF e classificados em sequência por Regressão Logística.

4.1. Geração de *Embeddings*

A geração de *embeddings* para os modelos LSTM e BERT envolveu os seguintes passos:

1. Cada palavra do *corpus* foi indexada por um número inteiro único (o tamanho do dicionário considerando todos os gêneros foi de 78307 palavras);
2. Cada letra de música foi associada a um vetor de inteiros, cujas componentes foram definidas pelos índices (em ordem) de suas palavras;
3. Cada vetor foi ampliado com zeros até que se alcançasse a dimensão daquele correspondente à letra com a maior quantidade (1996) de palavras, no caso do LSTM, e truncado ou ampliado com zeros até 512 palavras, no caso do BERT;
4. Cada vetor resultante foi mapeado num vetor real denso de 100 e 768 dimensões no caso dos modelos LSTM e BERT, respectivamente. No primeiro, por restrições computacionais. No segundo, por ser um valor comumente utilizado na literatura.

Cabe destacar que tais procedimentos foram realizados por códigos próprios para os modelos LSTM, e pela versão pré-treinada *BertTokenizerFast*, para o modelo BERT. Com relação às classificações por Regressão Logística, os *embeddings* gerados pelo BERT foram diretamente obtidos através do modelo pré-treinado através de sua versão *SentenceTransformer*. A abordagem TF-IDF, para uma melhor comparação dos resultados, considerou as 768 palavras mais frequentes das músicas, para todos os gêneros musicais.

5. Resultados

Este trabalho foi realizado utilizando a linguagem de programação *Python* e as bibliotecas de código aberto *Pandas* [Wes McKinney 2010], *Scikit-learn* [Pedregosa et al. 2011], *TensorFlow* [Abadi et al. 2016] e *Keras* [Chollet et al. 2015]. Os modelos foram treinados e avaliados através da plataforma “Google Colab”, considerando uma GPU “Tesla

T4” de 12GB. O desenvolvimento e a avaliação dos modelos consideraram a estratégia de *hold-out* [Japkowicz and Shah 2011], conforme as divisões explicitadas na Subseção 3.2.

Os hiperparâmetros dos modelos foram sintonizados de maneira gulosa e consideraram os valores presentes na Tabela 4. Tal sintonia começou pela LSTM padrão, assumindo $BS = 32$, por ser o valor intermediário. A busca pelo melhor número de épocas resultou em $NE = 5$, por *early stop*. Em seguida, fixando-se este valor, buscou-se o melhor tamanho do lote, dado por $BS = 8$. Logo após, passou-se à LSTM com *Dropout* e, mantidos os demais hiperparâmetros, os melhores resultados foram obtidos com $TD = 0,5$. Por fim, a BiLSTM com *Dropout* considerou os valores sintonizados anteriormente. Com relação ao BERT, a sintonia começou pelo número de épocas, fixando-se a taxa de aprendizado em $LR = 2e - 5$. A partir deste ponto, encontrou-se $NE = 3$ por *early stop*. Em seguida, de posse deste valor, buscou-se a melhor taxa de aprendizado, resultando em $LR = 5e - 5$. Para os classificadores por Regressão Logística, todos os hiperparâmetros adotados consideraram os valores padrões da biblioteca [Pedregosa et al. 2011].

Tabela 4. Valores avaliados durante a sintonia de hiperparâmetros para as redes do tipo LSTM e para o modelo BERT

Hiperparâmetro	Valores para LSTM	Valores para BERT
Número de épocas (NE)	2, 5, 10	2, 3, 4
Tamanho do lote (BS)	8, 16, 32, 64, 128	16
Taxa de <i>Dropout</i> (TD)	0,1, 0,25, 0,5	Não se aplica
Taxa de aprendizado (LR)	Não se aplica	2e-5, 5e-5, 1,25e-4

Para todos os modelos, a função-custo utilizada foi a “Entropia Cruzada Categórica” [Goodfellow et al. 2016] e o método de otimização empregado foi o “Adam” [Kingma and Ba 2015]. Ademais, como a base é balanceada e visando uma comparação mais direta com outros trabalhos, a figura de mérito considerada foi a acurácia.

A Tabela 5 sintetiza os resultados obtidos. Cabe observar que as duas abordagens baseadas em BERT apresentaram um desempenho melhor do que as baseadas em LSTM. Tal fato era em parte esperado, pois o BERT é um modelo pré-treinado e mais complexo, portanto mais hábil para a captura de nuances semânticas no conteúdo das letras. Ademais, em relação aos modelos LSTM, a técnica de *Dropout* combinada com a abordagem bidirecional resultou no modelo de maior acurácia dentre os desta categoria: 52,4%. Numericamente, os ganhos da estratégia de *Dropout* ($\sim 1,6\%$) se mostraram bem superiores aos associados ao uso de camadas bidirecionais ($\sim 0,6\%$). Cabe ainda ressaltar que o tempo de treinamento do modelo bidirecional foi praticamente o dobro do unidirecional. A abordagem BERT com ajuste fino atingiu a maior acurácia: 61,6%, valor similar ao percentual de acertos obtido por pessoas ao ouvir amostras de áudio com uma duração de 3 segundos (70%) [Gjerdingen and Perrott 2008] e, com relação à classificação a partir do conteúdo de letras em português, superior aos 50% de acertos reportados em [Guimarães et al. 2020], mesmo que esse último tenha considerado uma menor quantidade (7) de gêneros.

As acurácias por gênero foram 58,4% (forró), 78,4% (funk), 89,7% (gospel), 47,2% (MPB), 63,7% (pagode), 40,5% (pop-rock), 58,3% (rock), 59,1% (romântico), 55,3% (samba) e 64,4% (sertanejo). Assim, o modelo se mostrou mais acurado para o

Tabela 5. Resultados obtidos pelos modelos

Abordagem	Melhor configuração	Tempo de treino (minutos)	Acurácia no teste (%)
BERT (com ajuste fino)	$NE = 3$ $LR = 5e - 5$	147, 4	61, 6
BERT + Regressão Logística	(Sem ajuste fino)	5, 5	54, 5
BiLSTM com Dropout	$NE = 5, BS = 8$ $TD = 0, 5$	53, 3	52, 4
LSTM com Dropout	$NE = 5, BS = 8$ $TD = 0, 5$	28, 7	51, 8
LSTM	$NE = 5$ $BS = 8$	25, 9	50, 2
TF-IDF + Regressão Logística	(Sem ajuste fino)	0, 2	47, 3

gênero gospel, face ao seu vocabulário particular. Por sua vez, os maiores erros, 37,9% e 17,3%, se referem às letras de pop-rock classificadas como rock e vice-versa, respectivamente, o que está provavelmente associado à similaridade dos vocabulários, conforme as Tabelas 2 e 3. Outra medida relevante é a razão entre o total de canções corretamente classificadas em cada gênero pelo total de canções classificadas como pertencentes a este gênero, que foram 58,1% (forró), 87,4% (funk), 84,8% (gospel), 58,2% (MPB), 60,4% (pagode), 54,3% (pop-rock), 43,8% (rock), 46,8% (romântico), 63,4% (samba) e 66,6% (sertanejo). Convém notar que o gospel e o rock obtiveram o segundo maior e o menor entre todos os valores, respectivamente, o que está em acordo com a discussão ao final da Seção 3.

6. Conclusão

Este trabalho propôs a criação de uma base de dados balanceada, considerando letras de canções em português distribuídas em dez estilos musicais, e avaliou métodos de Aprendizado Profundo para automatizar sua classificação. O modelo BERT superou os demais, tendo sido atingida uma acurácia de 61,6%, valor próximo ao desempenho humano no reconhecimento de amostras de áudio com uma duração de 3 segundos.

Como trabalhos futuros, espera-se refinar os modelos, explorando a sintonia de outros hiperparâmetros e a pluralidade de estilos por artista, visto que a rotulagem pelo seu gênero mais frequente é uma das limitações do estudo presente. Ademais, almeja-se utilizar técnicas de validação cruzada para uma avaliação estatística mais rigorosa dos resultados. Considera-se também avaliar o modelo BERTimbau *large*, bem como modelos baseados na fusão de atributos derivados das letras e dos sinais de áudio das músicas.

7. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: A system for Large-Scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA. USENIX Association.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly.
- Bonds, M. E. (2018). *Listen to This*. Pearson, 4th edition.
- Chollet, F. et al. (2015). Keras. <https://keras.io>.
- da Silva Muniz, V. H. and de Oliveira e Souza Filho, J. B. (2023). Robust hand-crafted features for music genre classification. *Neural Computing and Applications*, 35(13):9335–9348.
- de Araújo Lima, R., de Sousa, R. C. C., Lopes, H., and Barbosa, S. D. J. (2020). Brazilian lyrics-based music genre classification using a blstm network. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I*, page 525–534. Springer-Verlag.
- de Oliveira, M. B. (2023). Conjunto de dados para classificação de gêneros a partir de letras de músicas em português. GitHub. <https://github.com/oliveiraamaatheus/Conjunto-de-dados-para-classificacao-de-generos-a-partir-de-letras-de-musicas-em-portugues> [Accessed: 2023-08-11].
- 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.
- Gjerdingen, R. and Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, 37(2):93–100.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Guimarães, P., Froes, J., Costa, D., and Freitas, L. (2020). A comparison of identification methods of Brazilian music styles by lyrics. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 61–63, Seattle, USA. Association for Computational Linguistics.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- HuggingFace (2018). Bert for sequence classification. Hugging Face Transformers Documentation. <https://huggingface.co/docs/transformers/mod>

- el_doc/bert#transformers.BertForSequenceClassification [Accessed: 2023-06-19].
- Japkowicz, N. and Shah, M. (2011). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, USA.
- Jeong, I. and Lee, K. (2016). Learning temporal features using a deep neural network and its application to music genre classification. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States*, pages 434–440.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Neisse, A. (2022). Song lyrics from 79 musical genres. Kaggle. <https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres> [Accessed: 2023-06-19].
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Pimenta, M. F. and Pugliesi, J. B. (2022). Reconhecimento de gêneros musicais com técnicas de aprendizagem de máquina supervisionada. *Revista Eletrônica de Computação Aplicada*, 3(1):23–46.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.
- 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*, page 403–417.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Vicente, D. (2022). Nem pagode, nem sertanejo: pisadinha faz o brasil dançar na pandemia. <https://www.exame.com.br/bussola/nem-pagode-nem-sertanejo-pisadinha-faz-o-brasil-dancar-na-pandemia> [Accessed: 2023-06-19].
- Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Sexismo no Brasil: análise de um *Word Embedding* por meio de testes baseados em associação implícita

Fernanda Tiemi de S. Taso¹, Valéria Q. Reis^{1,2}, Fábio V. Martinez¹

¹Faculdade de Computação, Universidade Federal de Mato Grosso do Sul,
Campo Grande, MS, Brasil

²Institute of Information Systems, Leuphana University,
Lüneburg, NS, Germany

{tiemi.taso, valeria.reis, fabio.martinez}@ufms.br

Abstract. *This work reports experiments based on the Psychology Implicit Association Test to identify and quantify biases in a Word Embedding (WE) of the Portuguese language. For this, we use a GloVe model trained on an Internet corpus collection. The results show that several common sense and gender stereotypes can be found in WE. Within the context of professions, we note a historical sexism, since the identified bias often reflects the statistics of gender performance in occupation groups in Brazil. The results show discrimination similar to those of international studies and allow discussing the impact of the use of language models in our society.*

Resumo. *Este artigo relata experimentos baseados no Teste de Associação Implícita da Psicologia para identificar e quantificar vieses em um Word Embedding (WE) de língua portuguesa. Para isso, usamos um modelo GloVe treinado em uma coleção de corpus da Internet. Os resultados mostram que diversos estereótipos de senso comum e de gênero podem ser encontrados no WE. Dentro do contexto de profissões, notamos um sexismo histórico, pois frequentemente o viés identificado reflete as estatísticas da atuação dos gêneros em grupos de ocupação do Brasil. Os resultados evidenciam discriminações semelhantes aos de estudos internacionais e permitem discutir sobre o impacto do uso de modelos de linguagem em nossa sociedade.*

1. Introdução

A discriminação de gênero, também chamada sexismo ou viés de gênero, é extensamente analisada na área de Processamento de Linguagem Natural (PLN). [Sun et al. 2019] classificaram os tipos de discriminação de gênero no PLN em quatro categorias, sendo: (a) difamação ou uso cultural ou histórico de termos depreciativos; (b) estereotipagem, que intensifica estereótipos sociais já existentes; (c) reconhecimento, que se refere à desproporção ou imprecisão de certo algoritmo em alguma tarefa de reconhecimento e (d) sub-representação, que define a baixa representação de certos grupos. Todas as quatro categorias são encontradas em modelos de *Word Embeddings* (WE). Exemplos de reconhecimento podem ser encontrados em máquinas de tradução [Tatman 2017, Prates et al. 2020].

[Prates et al. 2020] mostraram que o Google Tradutor exibe uma tendência a traduzir frases como “X é engenheiro”, onde “X” é um pronome neutro de idiomas que

não possuem flexão de gênero como o Húngaro, para o pronome masculino no inglês. Traduções desse tipo tornam-se mais frequentes quando nas sentenças são utilizadas profissões com sub-representação de mulheres na sociedade. São os casos das áreas de Ciência, Tecnologia, Engenharia e Matemática (do inglês *STEM*). Similarmente, essa desproporcionalidade de representação pode ser vista em modelos de linguagem onde a probabilidade condicional para profissões consideradas femininas ou masculinas são maiores para seus respectivos pronomes, amplificando padrões sexistas.

Ademais, [Suresh and Guttag 2021] demonstraram que modelos de WE podem apresentar viés histórico, o qual surge quando os sistemas produzem resultados prejudiciais e discriminatórios, apesar das medições e amostras nos dados terem sido feitas corretamente, refletindo os dados do mundo real. Em [Caliskan et al. 2017], os autores mostraram que modelos de WE conseguem captar relações implícitas de gênero, assim como ocorre nas respostas de participantes do Teste de Associação Implícita (IAT, do inglês) da Psicologia [Greenwald et al. 1998].

O teste IAT segue um paradigma de tempo de reação no qual os participantes são encorajados a classificar palavras rapidamente e o tempo de resposta observado quantifica a saída do teste. A reprodução do IAT com WE e PLN usa os mesmos atributos e palavras-alvo do trabalho original. No entanto, em vez de usar o tempo de resposta para a associação de palavras, usam a similaridade entre os vetores que as representam.

Além dos trabalhos que avaliam sexismo na língua inglesa usando PLN e modelos de WE, diversos trabalhos relacionados têm sido propostos em outras línguas tais como a chinesa [Chen et al. 2022, Li et al. 2022, Jiang et al. 2023, Qin et al. 2023], espanhola [Torres Berrú et al. 2023], alemã [Wagner and Zarriß 2022], filipina [Gamboa and Justina Estuar 2023], línguas africanas [Wairagala et al. 2022] e línguas indígenas [Hansal et al. 2022].

Carecem estudos sobre sexismo na língua portuguesa, mesmo com a disponibilização de diversos modelos de WE por [Hartmann et al. 2017] ainda em 2017. Assim, inspirado no trabalho de [Caliskan et al. 2017] e dando continuidade ao trabalho iniciado em [Taso et al. 2023], este artigo tem como objetivo verificar a existência de vieses de gênero por estereotipagem e sub-representação, utilizando metodologias similares às que foram empregadas em [Caliskan et al. 2017, Greenwald et al. 1998], mas utilizando um dos modelos de WE criados por [Hartmann et al. 2017]. Deve-se ainda analisar a relação dos vieses sexistas encontrados em profissões tradicionais com a proporção de mulheres no mercado de trabalho nacional.

Como resultado, este trabalho valida uma metodologia de identificação de vieses para WE da língua portuguesa, apresentando associações estereotipadas e, no caso da área profissional, sua relação com dados do mundo real. Adicionalmente, também abre caminho para o uso da metodologia no diagnóstico de outros tipos de discriminação.

2. Testes de associação implícita

O Teste de Associação Implícita (IAT) é um instrumento da Psicologia utilizado para quantificar o posicionamento de pessoas de maneira indireta, tal como o nome sugere. O seu uso é indicado em pesquisas onde os participantes não devem ou não querem expressar suas opiniões, mas as evidenciam ao associarem em tempos muito distintos pares de conceitos que consideram similares ou opostos [Greenwald et al. 1998].

Os criadores do IAT usaram o teste para comprovar alguns vieses humanos. Durante os experimentos, os participantes deviam associar dois conceitos, tais como flores e insetos, a um atributo, tal como agradável. Notaram-se tempos de resposta menores quando os envolvidos tiveram que classificar, por exemplo, um tipo de flor como agradável e um tipo de inseto como desagradável do que quando foram solicitados a fazer a classificação com os objetos trocados (flores como desagradáveis e insetos como agradáveis). O fato de um emparelhamento ser mais rápido indicaria que as partes envolvidas estão relacionadas no cognitivo dos indivíduos. Essa premissa motivou o uso do IAT para identificar e quantificar preconceitos étnicos e de estereótipo [Kiefer and Sekaquaptewa 2007, Nosek BA 2002].

2.1. WEAT

Word Embedding Association Test (WEAT) é um método variante do IAT, proposto para o diagnóstico de vieses em WE [Caliskan et al. 2017]. O WEAT assume que a similaridade por cosseno, métrica frequentemente utilizada para medir a semelhança semântica entre palavras representadas no espaço vetorial, é análoga ao tempo de reação do IAT, isto é, quanto menor o tempo de decisão, maior a proximidade semântica.

Em cada teste WEAT há dois conjuntos de palavras-alvo e dois conjuntos de atributos. Verifica-se então se o primeiro conjunto de palavras-alvo está mais associado ao primeiro conjunto de atributos e se o segundo conjunto de palavras-alvo está mais associado ao segundo conjunto de atributos. A hipótese nula é que não existe diferença de similaridade entre os conjuntos e seus respectivos atributos. O valor- p é utilizado para testar esta hipótese por meio do teste de permutação, e verificar a possibilidade de rejeitar a hipótese nula, ou seja, quanto menor o valor- p , maior a chance de rejeição. O valor- p de 10^{-2} foi mantido para rejeitar a hipótese assim como sugerido por [Caliskan et al. 2017].

Mais formalmente, considere o conjunto $W = X \cup Y$, onde X e Y são conjuntos-alvo de uma associação. Considere A e B seus respectivos conjuntos de atributos. A diferença entre as médias (μ) de similaridade entre uma palavra-alvo w , onde $w \in W$, e os conjuntos de atributos A e B é dada pela Equação 1:

$$s(w, A, B) = \mu_{a \in A} \{\cos(\vec{w}, \vec{a})\} - \mu_{b \in B} \{\cos(\vec{w}, \vec{b})\}. \quad (1)$$

Para exemplificar a Equação 1, considere $X = \{\text{rosa, azaléia, orquídea}\}$ o conjunto de palavras de flores, $Y = \{\text{formiga, pulga, mosca}\}$ o conjunto de insetos, $A = \{\text{paz, paraíso, arco-íris}\}$ o conjunto de palavras que denotam o conceito de agradável e $B = \{\text{fedor, veneno, agonia}\}$ o conjunto de palavras que denotam o conceito desagradável. A equação para $w = \text{rosa}$ pode ser lida da seguinte forma: $s(w, A, B)$ representa a média do cosseno entre *rosa* e todos os conceitos agradáveis de A menos a média do cosseno entre *rosa* e todos os conceitos desagradáveis de B . A mesma lógica pode ser utilizada para $w \in Y$.

O *effect size* d , medida que determina o tamanho da significância entre os conjuntos-alvos e seus respectivos atributos, é dado pela Equação 2, onde σ denota o desvio padrão:

$$d = \frac{\mu_{x \in X} \{s(x, A, B)\} - \mu_{y \in Y} \{s(y, A, B)\}}{\sigma_{w \in X \cup Y} \{s(w, A, B)\}}. \quad (2)$$

A medida da associação diferencial de dois conjuntos de palavras-alvo com os atributos é dada pela Equação 3:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B). \quad (3)$$

Por fim, o valor- p , medida estatística para determinar a probabilidade da significância do valor d para os conjuntos e seus atributos é dado pela Equação 4. Nela, os valores $\{(X_i, Y_i)\}_i$ representam a união dos conjuntos X e Y distribuídos aleatoriamente. A equação retorna o valor- p aproximado utilizando dez mil iterações com função de distribuição normal:

$$p = \Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]. \quad (4)$$

No estudo de [Caliskan et al. 2017] são usados os mesmos conjuntos de palavras adotados em [Greenwald et al. 1998] para a análise de um modelo GloVe treinado em textos encontrados na rede mundial de computadores. Naquele trabalho, todos os vieses linguísticos descritos no artigo original são também identificados, incluindo preconceitos raciais e de gênero. Por esse motivo, os proponentes do WEAT argumentaram que as associações nos vetores de representação dos WE não poderiam existir por casualidade e que seriam reflexo da perspectiva cultural da população.

A relevância acadêmica do trabalho de [Caliskan et al. 2017], a boa documentação dos experimentos realizados e a consolidação das métricas em trabalhos posteriores tornaram oportuna a validação da metodologia em modelos ainda pouco estudados, como de WE em português.

2.2. WEFAT

Word Embedding Factual Association Test (WEFAT) é um teste de associação implícita também proposto por [Caliskan et al. 2017]. Ele busca extrair informações empíricas sobre o mundo dentro de modelos de WE. Para isso, consideram-se, assim como no WEAT, um conjunto de palavras-alvo W e dois conjuntos de atributos A e B como definidos pela Equação 2.

Observe que o WEAT é usado para verificar se existem diferenças entre conjuntos de palavras-alvo em termos de sua similaridade relativa com conjuntos de atributo. Com o WEFAT, uma propriedade factual que pode ser valorada é associada a cada palavra-alvo e se deseja testar se os vetores correspondentes às palavras-alvo incorporam o conhecimento dessa propriedade, isto é, se é possível extrair ou prever a propriedade dado o vetor. Assim, o valor do WEFAT é utilizado para verificar a correlação com as informações do mundo real que foram informadas.

[Caliskan et al. 2017] utilizaram o WEFAT para mostrar que a representação de profissões em WE de língua inglesa embutem conhecimento sobre a composição da força de trabalho em ocupações nos Estados Unidos, ou seja, os estereótipos de profissão encontrados no WEAT apresentavam alta correlação com a proporção de mulheres atuantes naquela atividade. Neste trabalho, usaremos o teste para o mesmo objetivo dentro do contexto brasileiro e da língua portuguesa.

3. Procedimentos metodológicos

A metodologia seguida neste trabalho é consolidada na literatura no contexto americano e da língua inglesa. Para realizar experimentos no cenário brasileiro, foi preciso escolher um modelo de WE em português que já tivesse sido validado pela comunidade de PLN e contivesse um número suficiente de *tokens* para validar palavras nos mais variados escopos. Além disso, foram necessárias adaptações nas métricas adotadas para contemplar a marcação de gênero nas palavras.

3.1. Escolha do modelo

O modelo GloVe com 300 dimensões, avaliado por [Hartmann et al. 2017], demonstrou ter um bom desempenho em tarefas de analogias de sintaxe e semântica. Ele é formado pela junção de diferentes corpus, em português brasileiro e europeu, tendo mais de 1,2 bilhão de tokens e sendo amplamente utilizado para diversos tipos de aplicações [Grave et al. 2018, Fortuna et al. 2019, Garcia and Berton 2021, Silva et al. 2020]. Diante de seu bom resultado e uso no trabalho de [Caliskan et al. 2017], o modelo GloVe foi escolhido para ser usado na análise de discriminação de gênero em português.

3.2. Associações

Seis associações são analisadas: *Flores vs Insetos*, *Instrumentos vs Armas*, *Carreira vs Família*, *Matemática vs Artes*, *Ciência vs Artes*, e *Atuações Femininas vs Atuações Masculinas*. As duas primeiras fazem parte do grupo de associações universalmente aceitas como agradáveis ou desagradáveis. Elas servem principalmente para validar o WEAT em assuntos neutros, sobre os quais não há nenhuma questão social a ser discutida. As demais associações contêm supostos vieses de gênero e empregam os seguintes grupos de atributos:

- **Termos Femininos:** feminino, mulher, menina, irmã, ela, dela, delas, filha;
- **Termos Masculinos:** masculino, homem, menino, irmão, ele, dele, deles, filho.

Todas as associações, exceto as relacionadas a carreiras e áreas de atuação profissional, foram traduzidas de [Caliskan et al. 2017] para a língua portuguesa.

Para que o WEAT verificasse a existência de viés em profissões, foram construídos dois conjuntos de áreas de atuação. *Atuações Femininas* e *Atuações Masculinas* referem-se, respectivamente, às áreas de atuação com maior e menor proporção de mulheres no mercado de trabalho brasileiro de acordo com pesquisas de órgãos oficiais do Brasil¹. Esses conjuntos apresentam a seguinte formação²:

- **Atuações Femininas:** culinária, artes, educação, psicologia, pedagogia, enfermagem, assistência, estética, limpeza, farmácia, jornalismo, biblioteconomia, gastronomia, comunicação, literatura, sociologia, antropologia, nutrição, fisioterapia, música;
- **Atuações Masculinas:** atletismo, pesca, mecânica, comércio, indústria, agropecuária, ciência, economia, engenharia, física, medicina, diretoria, construção, administração, biologia, polícia, gerência, aviação, computação, direito.

¹ <https://bit.ly/3XjhZnw>, <https://bit.ly/46dBzp5>

² Os dados e códigos utilizados nos experimentos estão disponíveis em <https://github.com/nandayot/WEAT-WEFAT>.

3.3. Grupos ocupacionais

Para os experimentos com o WEFAT, foi criado um conjunto com 104 profissões cadastradas na Classificação Brasileira de Ocupações (CBO) agrupadas em 35 categorias ocupacionais de acordo com a similaridade de atuação. Paralelamente, obteve-se a proporção de mulheres atuantes em cada grupo, de acordo com dados do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2018 e outras instituições. Um exemplo de grupo ocupacional é o “Especialistas em métodos pedagógicos”, o qual apresenta 88% de mão de obra feminina. Nele constam profissões tais como pedagoga, psicopedagoga, fonoaudióloga e educadora.

4. Resultados

Os resultados foram divididos em duas categorias, onde se discute primeiramente o teste WEAT e, em seguida, o teste WEFAT.

4.1. WEAT

Na Tabela 1 são apresentados os resultados dos testes WEAT. Os valores obtidos para d apontam para vieses em todos os grupos de associação, indicando que os conjuntos de palavras-alvo possuem significativas diferenças de similaridade de acordo com os diferentes grupos de atributos³. Portanto, para todos os resultados, o 1º conjunto de palavras-alvo está mais associado ao 1º conjunto de atributos assim como o 2º está mais associado ao 2º conjunto de atributos.

Tabela 1. Resultados do teste WEAT. “A vs D” representam Agradável vs Desagradável e “TM vs TF” Termos Masculinos vs Termos Femininos.

Palavras-alvo	Atributos	Resultado	
		d	p
Flores vs Insetos	A vs D	0,87	10^{-3}
Instrumentos vs Armas	A vs D	0,91	10^{-4}
Carreira vs Família	TM vs TF	1,62	10^{-4}
Matemática vs Artes	TM vs TF	1,38	10^{-3}
Ciência vs Artes	TM vs TF	0,86	10^{-2}
Atuações Masculinas vs Atuações Femininas	TM vs TF	0,93	10^{-3}

Foi possível identificar vieses universais nas associações *Flores vs Insetos* e *Instrumentos vs Armas*. Dessa maneira, o WEAT está validado para relações assumidamente fortes, onde não há necessidade de discussão sobre posicionamentos, e abre espaço para a análise de associações que evidenciam discriminação de gênero.

Os grupos *Carreira vs Família* e *Matemática vs Artes* foram os que obtiveram os maiores *effect size* (d), com valores maiores do que os obtidos para os grupos de vieses universais. Esse resultado demonstra a existência de profundos estereótipos de gênero no campo de ocupações, com uma grande força de associações entre os conceitos e seus atributos. Assim, palavras tais como *executivo* e *carreira* estariam mais associadas a termos masculinos e palavras tais como *casa* e *filhos*, a termos femininos.

³Segundo [Caliskan et al. 2017], valores de d maiores que 0,8 indicam grande diferença de associação.

Ainda no contexto de profissões, notou-se uma clara associação das *Atuações Masculinas* com termos masculinos e das *Atuações Femininas* com termos femininos. Todos os vieses de gênero identificados colaboram na perpetuação de discriminação, principalmente contra as mulheres que, na sociedade atual, ainda são associadas a papéis relacionados à família, artes ou a profissões estereotipadas que envolvem, em sua maioria, cuidados e educação.

Todos os grupos obtiveram valores- p suficientes para refutar a hipótese de que não há diferença de associações entre os conjuntos-alvos. Dessa maneira, conclui-se que no WE utilizado identificam-se relações inquestionáveis, assim como associações carregadas de estereótipos de gênero. Os resultados vão ao encontro dos relatados em [Caliskan et al. 2017].

4.2. WEFAT

O teste WEFAT foi alterado para satisfazer as particularidades linguísticas da língua portuguesa, na concordância de gênero das palavras. Considere $\vec{p}_f$ e $\vec{p}_m$ palavras de profissões com flexão de gênero feminino e masculino, respectivamente (ex: advogada-advogado) e A e B os conjuntos de palavras que denotam os respectivos gêneros (Termos Femininos e Termos Masculinos, respectivamente). A nova fórmula para o cálculo da diferença entre as médias de similaridade entre palavras-alvo e atributos é dada pela Equação 5:

$$s(\vec{p}_f, \vec{p}_m, A, B) = \frac{\mu_{a \in A} \{\cos(\vec{p}_f, \vec{a})\} - \mu_{b \in B} \{\cos(\vec{p}_m, \vec{b})\}}{\sigma_{x \in A \cup B, w \in F \cup M} \{\cos(\vec{w}, \vec{x})\}}. \quad (5)$$

A Figura 1 ilustra a correlação entre a força de associação média de cada grupo ocupacional com a proporção de mulheres atuantes nele. O coeficiente de Pearson obtido foi de 0,86.

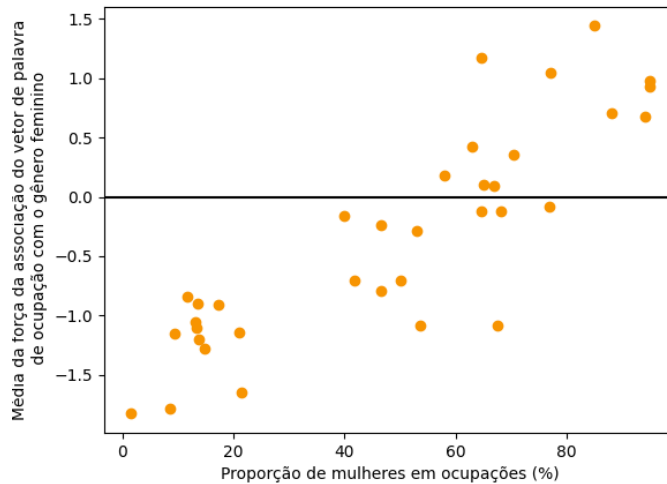


Figura 1. Relação entre a proporção de mulheres em ocupações no mercado de trabalho e a média de valores do teste WEFAT para ocupações com termos femininos e masculinos. Coeficiente de correlação de Pearson $p = 0,86$.

Observa-se que grande parte dos valores WEFAT estão abaixo do eixo x , indicando baixa força de associação com os termos femininos. De fato, muitas profissões, apesar de possuírem forte atuação de mulheres, possuem *embeddings* fracamente associados com termos femininos, o que pode indicar baixa co-ocorrência dessas palavras com estes termos dentro do modelo. A análise das frequências de palavras em *embeddings* feita por [Caliskan et al. 2022] corrobora essa hipótese. Os autores quantificaram o viés de gênero em grupos de 100, 1000, 10000 e 100000 mil palavras mais frequentes no modelo de língua inglesa e puderam verificar que 77% das mil palavras mais frequentes estão mais associadas a termos masculinos do que femininos.

Nota-se que a replicação do WEFAT consegue captar características reais do mercado de trabalho feminino no Brasil com um nível significativamente positivo de correlação.

5. Conclusões

Considerando que a Inteligência Artificial compreende a linguagem humana a partir de textos do mundo real, espera-se que os modelos de aprendizado gerados apresentem preconceitos encontrados nas sociedades onde os corpus se originaram. Tratando dessa hipótese, [Caliskan et al. 2017] mostraram que vieses universais, raciais e de gênero existem em WE da língua inglesa treinados em corpus obtidos na Internet. Para isso, os autores validam uma metodologia que utiliza duas novas métricas, WEAT e WEFAT, baseadas no Teste de Associação Implícita da Psicologia.

Neste trabalho, estendemos a proposta de [Caliskan et al. 2017] para o contexto brasileiro. Foi utilizado um modelo de Aprendizado de Máquina puramente estatístico treinado em diversos corpora com textos de páginas da Internet. As análises incluíram associações neutras, tidas como universais, assim como associações de gênero, dentro do contexto de profissões.

Os resultados mostraram a existência dos mesmos vieses humanos identificados em [Caliskan et al. 2017]. Além disso, é possível identificar vieses históricos no ramo de profissões do Brasil. Assim, concluímos que o uso de similaridade por cosseno é uma boa aproximação para a associação implícita de conceitos também na língua portuguesa. Nosso trabalho é a continuação do estudo iniciado em [Taso et al. 2023], sendo ambos pioneiros na detecção de vieses de dados no contexto brasileiro.

Críticas sobre o uso de associações e pares de gênero existem e devem ser levadas em consideração nas análises e discussões, mas reconhecemos que outras alternativas que contrapõem as métricas utilizadas ainda não são unanimidade dentro da área [Ethayarajh et al. 2019, Gonen and Goldberg 2019, Zhang et al. 2020]. Possíveis soluções para a mitigação de estereótipos em WE devem ser pensadas para os próximos trabalhos. Também propomos analisar outros tipos de modelo, assim como tratar diferentes tipos de discriminação e traçar a interseccionalidade entre eles.

A contribuição deste trabalho extrapola os limites da Computação. Entendemos que a interdisciplinaridade deve ser utilizada para abrir o escopo sobre como o sexismo em PLN pode ser entendido por meio de estudos da Sociolinguística e Ciências Sociais [Blodgett et al. 2020]. Além disso, o estudo sobre como aplicações de PLN impactam as comunidades que as utilizam deve ser essencial para os objetivos de pesquisa.

Agradecimentos

O presente trabalho foi realizado com apoio da Universidade Federal de Mato Grosso do Sul e da Universidade Leuphana de Lüneburg.

Referências

- Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. *arXiv preprint arXiv:2005.14050*.
- Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., and Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proc. of AAAI/ACM AIES*, pages 156–170.
- Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Chen, X., Li, M., Yan, R., Gao, X., and Zhang, X. (2022). Unsupervised mitigating gender bias by character components: A case study of Chinese word embedding. In *Proc. of GeBNLP*, pages 121–128. ACL.
- Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.
- Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proc. of AWL*, pages 94–104.
- Gamboa, L. C. and Justina Estuar, M. R. (2023). Evaluating gender bias in pre-trained filipino fasttext embeddings. In *Procc. of ITIKD*, pages 1–7.
- Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Appl Soft Comput*, 101:107057.
- Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.
- Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *J Pers Soc Psychol*, 74(6):1464–80.
- Hansal, O., Le, N. T., and Sadat, F. (2022). Indigenous language revitalization and the dilemma of gender bias. In *Proc. of GeBNLP*, pages 244–254. ACL.
- Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.
- Jiang, T., Li, Y., Fu, S., and Chen, Y. (2023). Creating a Chinese gender lexicon for detecting gendered wording in job advertisements. *Inform Process Manag*, 60(5):103424.
- Kiefer, A. K. and Sekaquaptewa, D. (2007). Implicit stereotypes and women’s math performance: How implicit gender-math stereotypes influence women’s susceptibility to stereotype threat. *J Exp Soc Psychol*, 43(5):825–832.

- Li, J., Zhu, S., Liu, Y., and Liu, P. (2022). Analysis of gender bias in social perception and judgement using Chinese word embeddings. In *Proc. of GeBNLP*, pages 8–16. ACL.
- Nosek BA, Banaji MR, G. A. (2002). Math = male, me = female, therefore math not = me. *J Pers Soc Psychol*, 83(1):44–59.
- Prates, M. O., 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.
- Qin, C., Zhang, X., Zhou, C., and Liu, Y. (2023). An interactive method for measuring gender bias and evaluating bias in Chinese word embeddings. In Imane, H., editor, *Proc. of CVAA*, volume 12613, page 126130U.
- Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Syst Appl*, 146:113199.
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.
- Suresh, H. and Gutttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proc. of EAAMO*, volume 17, pages 1–9.
- Taso, F. T., Reis, V. Q., and Martinez, F. V. (2023). Discriminação algorítmica de gênero: Estudo de caso e análise no contexto brasileiro. In *Anais do WICS*, pages 13–25. SBC.
- Tatman, R. (2017). Gender and dialect bias in YouTube’s automatic captions. In *Proc. of EthNLP*, pages 53–59. ACL.
- Torres Berrú, Y., Batista, V., and Zhingre, L. (2023). A data mining approach to detecting bias and favoritism in public procurement. *Intell Autom Soft Co*, 36(3):3501–3516.
- Wagner, J. and Zarrieß, S. (2022). Do gender neutral affixes naturally reduce gender bias in static word embeddings? In *Proc. of KONVENS*, pages 88–97.
- Wairagala, E. P., Mukiibi, J., Tusubira, J. F., Babirye, C., Nakatumba-Nabende, J., Katumba, A., and Ssenkungu, I. (2022). Gender bias evaluation in Luganda-English machine translation. In *Proc. of AMTA*, pages 274–286. AMTA.
- Zhang, H., Sneyd, A., and Stevenson, M. (2020). Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. *arXiv preprint arXiv:2010.02847*.

Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo “Universal Dependencies”

Emanuel Huber Silva^{1,3,4}, Thiago Alexandre Salgueiro Pardo¹, Norton Trevisan Roman²

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

²Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (USP)

³Centro de Inovação CESAR

⁴Departamento de Engenharia da Computação - Facens

emanuel.huber@usp.br, taspardo@icmc.usp.br, norton@usp.br

Abstract. *Part of speech tagging is a process that seeks to identify the grammatical classes of words and symbols (tokens) in a sentence. For Brazilian Portuguese, there is a variety of approaches using corpora of the journalistic genre with different tagsets. In this paper, we present results better than the current state of the art, investigating tagging methods and evaluating their ability to perform multi-genre analysis in corpora of journalistic, academic and user-generated content genres. To do so, we use the Universal Dependencies model. Finally, we present a qualitative assessment of the systematic tagging errors made in the process.*

Resumo. *A etiquetagem morfossintática é um processo que busca identificar as classes gramaticais de palavras e símbolos (tokens) em uma sentença. Para o português brasileiro, há uma variedade de trabalhos utilizando corpora de gênero jornalístico com diferentes conjuntos de etiquetas. Neste artigo, apresentamos resultados que superam o estado da arte atual, investigando métodos de etiquetagem e avaliando sua capacidade de análise multigênero em corpora dos gêneros jornalístico, acadêmico e de “user-generated content”. Para tanto, usamos o modelo “Universal Dependencies”. Por fim, apresentamos uma avaliação qualitativa dos erros sistemáticos cometidos pelo modelo.*

1. Introdução

A área de Processamento de Línguas Naturais (PLN) busca automatizar tarefas que envolvam a interpretação e a geração de língua natural [Jurafsky e Martin 2009]. Em várias dessas tarefas, faz-se necessário utilizar características linguísticas dos documentos, como as classes gramaticais de palavras e símbolos (ou etiquetas morfossintáticas dos *tokens* – do inglês, *part of speech tags*) e a estruturação sintática das sentenças.

Apesar da dominância atual das abordagens neurais e dos grandes modelos de língua, que na maioria das vezes processam textos em suas formas originais sem anotação linguística sofisticada, há muitas evidências da importância de informações linguísticas em PLN. Por exemplo, [Lin et al. 2021] combinam etiquetas morfossintáticas com representações vetoriais para aprimorar um analisador de opiniões baseado em aspectos. [Zhao et al. 2019], na frente de sumarização automática, demonstram a importância

de utilizar informações lexicais e de etiquetas morfossintáticas em conjunto com mecanismos de atenção. [Cabral et al. 2022], por sua vez, fazem uso desses conhecimentos no desenvolvimento de um sistema de extração de informação aberta para o português. [Garimella et al. 2019], em um estudo socio-linguístico, demonstram que há diferenças gramaticais em textos escritos por homens e mulheres.

Motivadas pela importância desse tipo de conhecimento em PLN, há várias iniciativas clássicas e mais recentes para o desenvolvimento de recursos e ferramentas relacionados para o processamento computacional da língua portuguesa. Pode-se citar, por exemplo, a amplamente conhecida Floresta Sintá(c)tica [Afonso et al. 2002] e o *treebank* Porttinari [Pardo et al. 2021], o léxico de língua geral PortiLexicon-UD [Lopes et al. 2022], o etiquetador morfossintático LX-Tagger [Branco e Silva 2004] e o etiquetador do estado da arte de [Fonseca et al. 2015] treinado com o corpus de referência Mac-Morpho [Aluísio et al. 2003], assim como o conhecido *parser* PALAVRAS [Bick 2000], entre muitas outras pesquisas relevantes.

Visando a contribuir nesta frente e avançar a fronteira do conhecimento, este artigo foca na tarefa de etiquetagem morfossintática para o português, mas trazendo ambições maiores. Por um lado, são investigados métodos variados e do estado da arte para conjuntos de dados de referência em português, avaliando-se a capacidade de análise multigênero dos métodos. Objetiva-se, com isso, o desenvolvimento de um etiquetador de alta acurácia e de amplo uso, possibilitando o desenvolvimento de aplicações de PLN mais robustas. Para tanto, utilizam-se os corpora Porttinari [Pardo et al. 2021], DANTES-tokens [Di Felippo et al. 2021] e PetroGold [Souza et al. 2021], dos gêneros jornalístico, gerado por usuário (do inglês, *User-Generated Content* - UGC) e acadêmico (do domínio de óleo e gás), respectivamente. Por outro lado, explora-se o modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021], de ampla aceitação, inclusive para o português [Rademaker et al. 2017]. Mostramos que nossos melhores resultados ultrapassam 99% de acurácia e que é possível produzir um etiquetador morfossintático multigênero de alta acurácia, superando o estado da arte. Mais do que isso, na análise qualitativa realizada, evidencia-se que muitos dos erros remanescentes são linguisticamente plausíveis.

O restante desse trabalho está organizado como segue. Na Seção 2, os trabalhos relacionados são sucintamente apresentados. Na Seção 3, os corpora utilizados são introduzidos. Os experimentos realizados e os resultados atingidos são relatados nas Seções 4 e 5. Por fim, a Seção 6 conclui esse trabalho.

2. Trabalhos relacionados

Há vários trabalhos em etiquetagem morfossintática para o português, dos quais destacamos alguns. [Fonseca et al. 2015] utilizam uma rede neural com representações vetoriais das palavras e atributos linguísticos adicionais (como capitalização e sufixos) para prever suas etiquetas. Os autores utilizam diferentes versões do corpus jornalístico Mac-Morpho [Aluísio et al. 2003], atingindo 97,57% de acurácia (ou seja, a proporção de *tokens* corretamente classificados). Utilizando o mesmo corpus, [de Sousa e Lopes 2019] avaliam as Redes Neurais Recorrentes (RNRs) bidirecionais com representações vetoriais em nível de palavra e caractere. Essa abordagem alcançou 97,36% de acurácia. [Domingues 2011] apresenta um etiquetador que utiliza o aprendizado baseado em transformações para os gêneros jornalístico e acadêmico. Foram utilizados um léxico

Tabela 1. Exemplos dos corpora selecionados

Corpus	Exemplo
Porttinari-base	Foram/AUX avaliados/VERB 5.281/NUM municípios/NOUN ./PUNCT ou/CCONJ 95/NUM %/SYM de/ADP o/DET total/NOUN de/ADP 5.569/NUM existentes/ADJ em/ADP o/DET Brasil/NOUN ./PUNCT
DANTEStocks	BBAS3/PROPN comprar/VERB por/ADP R\$/SYM 20,05/NUM indicado/VERB em/ADP 27/02/2014/NUM 10:41/NUM http://t.co/zJR3Eeyz9/SYM
PetroGold	Segundo/ADP Luiz/PROPN &/PROPN Silva/PROPN (/PUNCT 1995/NUM)/PUNCT estas/DET feições/NOUN definem/VERB a/DET maioria/NOUN de/ADP os/DET lineamentos/NOUN em/ADP mapas/NOUN magnéticos/ADJ ./PUNCT

para o tratamento de nomes próprios, regras manuais e a saída de outros dois etiquetadores disponíveis na literatura. Além do Mac-Morpho, o trabalho também utilizou o Bosque (que integra a Floresta Sintá(c)tica) para o gênero jornalístico. Para o gênero acadêmico, utilizou a Selva Científica (também parte da Floresta Sintá(c)tica). A avaliação apresentou acurácias de 98,06%, 98,30% e 98,07%, respectivamente. Outros trabalhos baseados em RNRs e com uso de diferentes representações vetoriais alcançaram alto desempenho no corpus Bosque. Destacam-se o UDPipe 2 [Straka 2018], com 96,37% de acurácia, o CNC SR [Heinzerling e Strube 2019], com 98,1%, e o Stanza [Qi et al. 2020], com 97,04%. Por fim, destaca-se o trabalho de [Bohnet et al. 2018], que utiliza a técnica de Meta-BILSTM, com a premissa de que o uso de diferentes representações vetoriais pode contribuir para o desempenho na tarefa. O modelo alcançou 98,11% de acurácia no corpus Bosque.

Os conjuntos de etiquetas morfossintáticas (*tagsets*) variam nos diferentes trabalhos. Os trabalhos mais recentes fazem uso do *tagset* do modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021], composto por 17 etiquetas. As classes abertas são representadas pelas etiquetas ADJ, ADV, INTJ, NOUN, PROPN e VERB; as classes fechadas são ADP, AUX, CCONJ, DET, NUM, PART, PRON e SCONJ; há também as etiquetas para outros casos, como PUNCT, SYM e X. O modelo UD já é adotado por mais de 100 línguas, contando com aproximadamente 200 *treebanks* catalogados. Esse modelo tem tido grande aceitação em função de sua proposta de “universalidade”, com aplicação para línguas tipologicamente diferentes, já tendo passado por algumas versões. Como comentado anteriormente, este trabalho também se filia ao modelo UD.

3. Corpora

Neste trabalho, foram utilizados três corpora de gêneros diferentes, anotados manualmente segundo o modelo UD. Para o gênero jornalístico, foi utilizada a porção “base” do *treebank* Porttinari [Pardo et al. 2021], com notícias do jornal Folha de São Paulo. A porção “base” é a semente com base na qual o restante do *treebank* foi anotado. Para o gênero de UGC, adotou-se o corpus DANTEStocks [Di Felippo et al. 2021], que contém *tweets* do mercado financeiro. Contemplando o gênero acadêmico, o corpus PetroGold [Souza et al. 2021] apresenta uma coletânea de textos da área de óleo e gás, provenientes de teses, dissertações e monografias. Na Tabela 1, para evidenciar os desafios da tarefa, é possível visualizar um exemplo manualmente anotado de sentença ou *tweet* de cada corpus (a etiqueta morfossintática é separada dos *tokens* pela barra).

A Tabela 2 mostra o total de sentenças e *tokens* de cada corpus. É possível observar que o corpus DANTEStocks tem uma quantidade menor de *tokens* quando comparado aos corpora Porttinari-base e PetroGold. Ressalta-se que os corpora DANTEStocks e Porttinari-base originalmente não possuem a divisão em conjuntos de treino, validação e teste. Dessa forma, para fins de avaliação e comparação justa entre métodos, foi realizada essa divisão com a amostragem aleatória, utilizando a proporção de 10% para validação e 20% para o conjunto de teste, resultando nos números mostrados na tabela.

Tabela 2. Estatísticas dos corpora utilizados

Corpus	Gênero	Treino	Validação	Teste	Sentenças	<i>tokens</i>
Porttinari-base	Jornalístico	5.894	585	1.668	8.420	168.400
DANTEStocks	UGC	2.833	413	802	4.048	81.048
PetroGold	Acadêmico	8.054	447	445	8.946	250.905

É interessante notar dois pontos adicionais sobre os corpora selecionados. Em primeiro lugar, eles contêm textos bastante diferentes entre si, tanto em gênero quanto domínio. Isso é importante para o teste que este artigo se propõe a fazer, de avaliar a capacidade multigênero dos métodos. Em segundo lugar, há outros corpora que são anotados com UD e disponibilizados publicamente, como o Bosque [Rademaker et al. 2017], o CINTIL [Branco et al. 2022] e o PUD (*Parallel Universal Dependencies*) [Zeman et al. 2017], mas que foram preteridos por não seguirem diretrizes de anotação similares e não conterem apenas textos em português brasileiro. Os três corpora selecionados, além de serem para o português brasileiro, fazem parte de um esforço nacional de estudo e uniformização de UD para o português¹. Dessa forma, há menos variáveis envolvidas nos experimentos realizados.

4. Experimentos

A experimentação foi dividida em duas etapas. A primeira consistiu em avaliar diferentes técnicas de etiquetagem no corpus jornalístico, o Porttinari-base. Em seguida, aplicou-se no contexto multigênero a técnica de melhor desempenho, considerando então os demais corpora. Essa estratégia visou a otimizar a sequência de testes necessários.

4.1. Técnicas de etiquetagem morfossintática

Foram selecionadas sete técnicas/modelos de etiquetagem morfossintática para a avaliação no corpus Porttinari-base, sendo esta seleção feita com base na representatividade e no desempenho dessas técnicas na literatura.

O primeiro modelo, UDPipe 2 [Straka 2018], foi avaliado com o tamanho de lotes (*batch size*) de 128 amostras, com um treinamento de 16 épocas, onde, nas primeiras 8 épocas, é utilizada a taxa de aprendizagem de 10^{-3} , e de 10^{-4} nas demais. Como modelo de língua, foi utilizado o BERTimbau [Souza et al. 2020].

O Stanza [Qi et al. 2020] possui um módulo de etiquetagem morfossintática que utiliza redes Bi-LSTM para a classificação. Para este modelo, foi utilizado o tamanho em lotes padrão de 5.000, taxa de aprendizagem de 10^{-3} e número máximo de atualizações de etapas de gradiente de 1.000.

¹<https://sites.google.com/icmc.usp.br/poetisa>

O terceiro modelo, Meta-BiLSTM [Bohnet et al. 2018], foi treinado com o tamanho de lotes de 40.000 para o modelo em nível de palavras e 80.000 para o modelo em nível de caracteres. A taxa de aprendizagem é de 2×10^{-3} , com 3 camadas ocultas com 400 neurônios cada. O modelo utiliza representações estáticas em nível de palavra, obtidas do Skip-gram do Word2Vec com dimensão 300 [Hartmann et al. 2017].

Outra técnica foi a CNCSR [Heinzerling e Strube 2019], que se baseia no uso de representações vetoriais em nível de palavra e caractere com rede Bi-LSTM. Foram utilizadas as representações em nível de caractere e subpalavra, sendo elas combinadas por meio de uma rede RNR meta. O modelo foi treinado com tamanho de lotes de 64, número de épocas mínimo de 50 e máximo de 1.000, taxa de aprendizagem de 10^{-4} , tamanho de vocabulário de 100.000 e taxa de *dropout* de 0, 2. O modelo em nível de caractere possui representação vetorial de tamanho 50 e camada oculta com 256 neurônios; os modelos de subpalavra e meta possuem o mesmo número de neurônios na camada oculta.

Além destes modelos, foram realizados experimentos com três diferentes modelos de língua em conjunto com etapas de ajuste fino. Dessa forma, são utilizadas as representações da primeira subpalavra de cada *token* da sentença de entrada para detectar a classe gramatical. Foram utilizados os modelos de língua BERTimbau [Souza et al. 2020], DeBERTa-v3 [He et al. 2021] e XLM-R [Conneau et al. 2020]. Para os três modelos, foram utilizados os seguintes hiper-parâmetros: máximo de 30 épocas, taxa de aprendizagem de 2×10^{-5} e *weight decay rate* de 0, 01, que é um parâmetro do otimizador AdamW [Loshchilov e Hutter 2019]. Os modelos BERTimbau e XLM-R utilizaram tamanho de lotes de 32 e, para o DeBERTa-v3, foi utilizado tamanho 16.

O procedimento experimental conta com a realização de 10 execuções² de treinamento no conjunto de treino do corpus Portinari-base, para, então, realizar a comparação entre os modelos e realização de testes de hipótese para identificar diferenças estatisticamente significativas na acurácia. O teste Anova [Fisher 1992] com *post hoc* de Tukey [Tukey 1949] foi selecionado para realizar esta avaliação. O teste Anova avalia se existem diferenças significativas entre as médias de dois ou mais grupos. Se identificada tal diferença, o teste de Tukey é aplicado para determinar quais os grupos que possuem médias significativamente distintas entre si, com correção para múltiplas testagens.

4.2. Resultados

A Tabela 3 apresenta os resultados da avaliação da etiquetagem morfosintática no corpus jornalístico. São apresentadas a acurácia média e a Medida-F Macro média das 10 execuções de experimentos para cada abordagem avaliada, além dos respectivos desvios padrões. É possível observar que os métodos baseados em RNRs possuem desempenho inferior aos métodos baseados em modelos de língua com ajuste fino, tanto em termos de acurácia quanto em Medida-F macro. Além disso, a abordagem com o BERTimbau possui o maior valor absoluto médio para acurácia e Medida-F Macro. Os modelos DeBERTa-v3 e XLM-R possuem valores próximos. As diferenças observadas com relação à acurácia foram significativas (Anova $Z(70, 69) \approx 890, p = 6e - 59$), com nível de confiança de 95%). Em análise par-a-par, as diferenças observadas foram significativas para todos os pares, exceto para BERTimbau $\times$ DeBERTa-v3 e XLM-R $\times$ DeBERTa-v3.

²Cada experimento utilizou o mesmo conjunto de treinamento, com variação na semente aleatória que é utilizada na inicialização dos pesos do modelo.

Tabela 3. Acurácia no corpus jornalístico Porttinari-base

Modelo	Abordagem	Acurácia média (%)	Medida-F macro média (%)
BERTimbau	Modelo de língua	99,07 ± 0,03	96,39 ± 0,32
DeBERTa-v3	Modelo de língua	99,02 ± 0,05	95,81 ± 0,39
XLM-R	Modelo de língua	99,00 ± 0,04	96,36 ± 0,42
Meta-BiLSTM	RNR	98,47 ± 0,06	94,89 ± 0,28
Udpipe 2	RNR	98,01 ± 0,03	93,13 ± 0,54
Stanza	RNR	98,22 ± 0,05	94,60 ± 0,27
CNCSR	RNR	98,10 ± 0,07	94,04 ± 0,30

Dado que não foi observada diferença significativa entre os métodos baseados nos modelos BERTimbau e DeBERTa-v3, o método baseado no BERTimbau foi selecionado para a próxima etapa de experimentação, devido a seu menor número de parâmetros. O método foi avaliado nos três corpora de gêneros diferentes (jornalístico, acadêmico e UGC), em que o experimento é constituído pelo treinamento do modelo em cada cenário de combinação dos corpora, seguido de sua avaliação separada em cada corpus individual. A [Tabela 4](#) exibe a acurácia média dos experimentos nos conjuntos de teste.

Tabela 4. Acurácia no contexto multigênero

Corpora de treinamento	Acurácia média (%)		
	Porttinari-base	DANTEStocks	PetroGold
Porttinari-base	99,07 ± 0,03	87,14 ± 0,60	96,46 ± 0,17
DANTEStocks	96,55 ± 0,23	97,98 ± 0,08	94,95 ± 0,20
PetroGold	96,99 ± 0,10	84,96 ± 0,46	98,93 ± 0,06
Porttinari-base + DANTEStocks	99,05 ± 0,04	97,91 ± 0,10	96,58 ± 0,16
Porttinari-base + PetroGold	98,95 ± 0,06	85,29 ± 0,34	98,85 ± 0,07
DANTEStocks + PetroGold	97,86 ± 0,06	97,99 ± 0,07	98,92 ± 0,05
Port.-base + DANTEStocks + PetroGold	99,00 ± 0,05	97,92 ± 0,13	98,89 ± 0,06

É possível observar que o cenário que obteve a maior acurácia média foi o cenário onde o modelo foi treinado apenas com dados do gênero alvo. Por exemplo, o melhor cenário para o corpus de gênero acadêmico foi o cenário em que o treinamento foi exclusivamente neste gênero. Contudo, estes modelos possuem acurácias mais baixas nos outros gêneros, por exemplo, o modelo treinado no corpus PetroGold com acurácia de 98,93% no gênero acadêmico possui acurácia de 84,96% no gênero UGC.

Também se pode notar maior discrepância entre os gêneros que seguem a norma culta da língua e o gênero UGC, que possui características linguísticas diferentes. Quando o cenário com o PetroGold é avaliado no gênero jornalístico, por exemplo, é possível observar uma acurácia de 96,99% (ou seja, há uma diferença relativamente pequena em relação ao melhor resultado para esse gênero). Já no gênero UGC, observa-se uma diferença maior em relação ao melhor modelo treinado no corpus DANTEStocks.

Em relação ao treinamento multigênero, é possível observar que o modelo treinado em todos os gêneros (última linha da tabela) alcançou desempenho similar aos modelos treinados isoladamente, sendo que a diferença entre as médias possui valor máximo de 0,067. Como esperado, essa diferença não foi estatisticamente significativa³.

³Anova $Z(70, 69) \approx 1107, p = 7e - 62$. Tukey: Multigênero vs Porttinari-base $Z \approx 0,7e - 4, p \approx$

Além da acurácia em nível de *tokens*, também foi calculada a acurácia em nível de sentença nos corpora, computando-se a porcentagem de sentenças que foram anotadas de forma completamente correta, obtendo-se os seguintes resultados médios nos corpora: Porttinari-base – 64,59%; DANTEStocks – 54,25%; PetroGold – 47,36%; Porttinari-base + DANTEStocks – 68,31%; Porttinari-base + PetroGold – 55,01%; DANTEStocks + PetroGold – 72,79%; Porttinari-base + DANTEStocks + PetroGold – 77,70%. Novamente, o cenário multigênero destaca-se. Aprofundando o estudo, na análise das sentenças com erros no cenário multigênero, é possível observar que: em 77% das sentenças, houve apenas 1 erro; em 18%, dois erros; em 4%, 3 ou 4 erros; o restante ($< 1\%$) tem 5 ou mais erros (que incluem casos de sentenças de estrutura incomum). Os resultados indicam um novo estado da arte para a língua portuguesa, além de demonstrarem que é possível ter um sistema multigênero robusto que possibilite o desenvolvimento de aplicações de PLN mais generalistas e que possam ser aplicados para textos variados.

Após a avaliação quantitativa, partiu-se para a avaliação qualitativa, essencial para entender a potencialidade real desse tipo de sistema e suas limitações. Partindo do modelo treinado no contexto multigênero, foi realizada a análise manual de erros (com o apoio de um linguista experiente), buscando-se encontrar erros ocorridos para cada etiqueta morfossintática. Aqui são reportados apenas os erros sistemáticos observados.

Com relação à etiqueta ADJ, no corpus Porttinari-base, foram encontrados 23 casos onde os *tokens* estavam na forma de particípio. Particípio é uma forma nominal do verbo e pode assumir as etiquetas ADJ, NOUN ou VERB, sendo um caso particularmente desafiador para a Linguística [Duran 2021]. Naturalmente, o mesmo tipo de erro é encontrado ao analisar os erros das etiquetas NOUN e VERB. Nos corpora DANTEStocks e PetroGold, foram encontradas 4 ocorrências em ambas as análises.

Para a etiqueta PROP, é possível identificar casos em que o modelo classificou como NOUN, consistindo em outra dificuldade conhecida da área. No corpus Porttinari-base, foram 10 ocorrências; no DANTEStocks, 21; e 8 ocorrências no PetroGold. Em especial, no DANTEStocks, foi observado que alguns *tweets* continham índices da bolsa de valores sendo classificados com a etiqueta X. No total, foram encontradas 30 ocorrências desse tipo. Esse corpus adotou a etiqueta X para índices da bolsa que não possuíam função linguística no *tweet* e, quando possuíam função, a etiqueta PROP deveria ser utilizada.

Finaliza-se com a etiqueta X, utilizada para casos a que outras etiquetas não podem ser associadas. No corpus Porttinari-base, todos os erros encontrados foram casos de estrangeirismos a que o modelo tentou associar uma classe gramatical diferente da etiqueta X. Este tipo de erro foi encontrado em 11 casos no corpus DANTEStocks e não ocorreu no corpus PetroGold. Esse é um erro considerado plausível, já que estrangeirismos poderiam ter outras etiquetas associados a eles.

É interessante observar que, caso esses erros relatados fossem computados como análises plausíveis no cálculo da acurácia, a acurácia geral do melhor modelo de etiquetagem se aproximaria dos 100%. Esses casos também podem servir de base para futuras discussões e eventuais aprimoramentos nos corpora anotados.

0,77, Multigênero vs DANTEStocks $Z \approx 0,7e-4$, $p \approx 0,99$, Multigênero vs PetroGold $Z \approx 0,4e-4$, $p \approx 0,99$ ao nível de confiança de 95%.

5. Experimentos adicionais: o corpus Mac-Morpho

Dada a relevância histórica do corpus Mac-Morpho [Aluísio et al. 2003] para a tarefa de etiquetagem morfossintática para o português, testamos nesse corpus a melhor técnica de etiquetagem observada no experimento anterior. O Mac-Morpho contém cerca de 1 milhão de palavras em português brasileiro, criado a partir de textos de jornais e revistas. A versão atual, Mac-Morpho v2 [Fonseca e Rosa 2013], conta com 23 etiquetas morfosintáticas de base e 7 complementares. Sendo assim, o conjunto de etiquetas é distinto do conjunto da UD. A contribuição desses experimentos adicionais reside, portanto, na avaliação da robustez da melhor técnica identificada em dados com um *tagset* diferente.

A Tabela 5 exibe as acurácias obtidas por trabalhos prévios da literatura e pelo etiquetador deste artigo baseado no modelo BERTimbau. É possível observar que o etiquetador deste trabalho obteve a maior acurácia, demonstrando sua robustez e avançando o estado da arte de etiquetagem para o corpus Mac-Morpho também.

Tabela 5. Acurácia para o corpus Mac-Morpho

Método	[Fonseca e Rosa 2013]	[de Sousa e Lopes 2019]	[Fonseca et al. 2015]	[Santos e Zadrozny 2014]	BERTimbau
Acurácia	96,48%	97,62%	97,31%	97,47%	98,36%

6. Considerações finais

Este trabalho avançou a fronteira do conhecimento e o estado da arte ao demonstrar a potencialidade multigênero de um método de etiquetagem morfossintática baseado em modelagem de língua e ao produzir resultados superiores ao estado da arte.

O melhor método observado, baseado no modelo BERTimbau, demonstrou uma boa capacidade de generalização nos gêneros abordados, mas pode ser interessante no futuro avaliá-lo ainda em outros gêneros e domínios a fim de confirmar tal robustez. Outro fator importante a ser considerado é o custo computacional desse etiquetador. Possuindo cerca de 110 milhões de parâmetros e complexidade quadrática no mecanismo de auto-atenção, o tempo de inferência é considerável. Pode ser interessante explorar técnicas de compressão de modelos para reduzir o tamanho e tempo de inferência.

Para reprodução dos resultados apresentados, o repositório⁴ de código é disponibilizado. Além disso, uma aplicação⁵ foi criada para que interessados possam utilizar o melhor etiquetador desenvolvido (no cenário multigênero ou não). Outras informações sobre este trabalho e sobre iniciativas relacionadas podem ser encontradas no portal web do projeto POeTiSA⁶.

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.

⁴<https://github.com/huberemmanuel/porttagger>

⁵<https://huggingface.co/spaces/Emanuel/porttagger>

⁶<https://sites.google.com/icmc.usp.br/poetisa/>

Referências

- Afonso, S., Bick, E., Haber, R., e Santos, D. (2002). Floresta sintá(c)tica: A treebank for Portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1698–1703, Las Palmas, Spain.
- Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., e Marquiasfável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In *6th international conference on Computational processing of the Portuguese language*, page 110–117, Faro, Portugal.
- Bick, E. (2000). *The Parsing System “Palavras”. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus.
- Bohnet, B., McDonald, R., Simões, G., Andor, D., Pitler, E., e Maynez, J. (2018). Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2642–2652, Melbourne, Australia.
- Branco, A. e Silva, J. (2004). Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 507–510, Lisbon, Portugal.
- Branco, A., Silva, J. R., Gomes, L., e António 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, Marseille, France.
- Cabral, B., Souza, M., e Claro, D. B. (2022). Portnoie: A neural framework for open information extraction for the portuguese language. In *Computational Processing of the Portuguese Language: 15th International Conference*, page 243–255, Berlin, Heidelberg.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., e 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.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., e Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47:255–308.
- de Sousa, R. C. C. e Lopes, H. (2019). Portuguese pos tagging using blstm without hand-crafted features. In Nyström, I., Hernández Heredia, Y., e Milián Núñez, V., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 120–130, Havana, Cuba.
- Di Felippo, A., Postali, C., Cereghatto, G., Gazana, L., Silva, E., Roman, N., e 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*, pages 335–343, Porto Alegre, RS, Brasil.
- Domingues, M. L. C. S. (2011). *Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o Português do Brasil*. PhD thesis, Universidade Federal do Pará, Belém, PA, Brasil.

- 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). Technical report, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, São Carlos, Brasil.
- Fisher, R. A. (1992). *Statistical Methods for Research Workers*. Springer New York.
- Fonseca, E. R., G Rosa, J. L., e Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21:1–7.
- Fonseca, E. R. e Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 1–10, Fortaleza, Brasil.
- Garimella, A., Banea, C., Hovy, D., e Mihalcea, R. (2019). Women’s syntactic resilience and men’s grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy.
- Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., e Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, Brasil.
- He, P., Gao, J., e Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543:1–19.
- Heinzerling, B. e Strube, M. (2019). Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy.
- Jurafsky, D. e Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.
- Lin, Y., Wang, C., Song, H., e Li, Y. (2021). Multi-head self-attention transformation networks for aspect-based sentiment analysis. *IEEE Access*, 9:8762–8770.
- Lopes, L., Duran, M., Fernandes, P., e Pardo, T. (2022). PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France.
- Loshchilov, I. e Hutter, F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, pages 1–19, Toulon, France.
- Pardo, T., Duran, M., Lopes, L., Felippo, A. D., Roman, N., e 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, Brasil.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., e 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*, pages 101–108, Online.
- Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., e de Paiva, V. (2017). Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 197–206, Pisa, Italy.
- Santos, C. D. e Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1818–1826, Beijing, China.
- Souza, E., Silveira, A., Cavalcanti, T., Castro, M., e 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, Brasil.
- Souza, F., Nogueira, R., e Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham.
- 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.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5:99–114.
- Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyy-salo, 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., Droганova, 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., Al-calde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., e 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.
- Zhao, F., Quan, B., Yang, J., Chen, J., Zhang, Y., e Wang, X. (2019). Document summarization using word and part-of-speech based on attention mechanism. *Journal of Physics: Conference Series*, 1168:32008.

Automated question answering via natural language sentence similarity: Achievements for Brazilian e-commerce platforms

Víctor Jesús Sotelo Chico¹, Luiz Zucchi^{1,2}, Daniel Ferragut^{1,2}, Rodrigo Caus²,
Victor Hochgreb de Freitas², Julio Cesar dos Reis¹

¹Institute of Computing – University of Campinas, Campinas, Brazil

²GoBots Company, Brazil

jreis@ic.unicamp.br, {rodrigo.caus,victor}@gobots.com.br

{victor.sot.c, luizeduardoaraujozucchi, danielpferragut}@gmail.com

Abstract. *Chatbots have become indispensable for answering e-commerce customer queries, which is crucial for selling products online. However, in Brazilian e-commerce, finding scalable solutions can be challenging. This article proposes an automatic question-answering system by replying to incoming questions with Frequently Asked Questions from stores. Our solution builds a store-specific database populated with question-answer pairs by generating the embedding of questions. We rank candidate questions using a neural network to retrieve known answers. Our solution was tested on data from South American e-commerce platforms in Portuguese and Spanish. The development approach achieved 97.75% of satisfaction with the given answers.*

1. Introduction

Customer service has become an essential part of e-commerce. Offering an excellent experience to clients through a virtual platform is the key to establishing trust. Customers spend hours researching products with thousands of choices available, and supporting eventual questions in real time about the product is vital to guarantee a sale.

E-commerce companies focus on automatizing this process with artificial intelligence solutions; some solutions involve finding the question's *intent* (the user's purpose) and its *entities* (relevant terms and objects in the query). This approach is hard to scale because answering a question requires knowing all possible intents and entities for that type of product. Furthermore, while large language models can provide answers for e-commerce stores, they are often private and too expensive for small businesses.

This article offers a solution for customer questions by comparing them to previously answered ones. If a past customer has asked a question about a product before, our solution uses the answer provided by customer service. The defined approach does not rely on detecting any *intent* or *entities* for a product or an incoming question in free text.

We face a challenging task in sentence similarity detection. First, it requires a coherent way to represent sentences so that questions presenting similar meanings also have similar numeric values. In this context, the main issue is to obtain models for our specific goal, which must address the informal Portuguese language spoken on e-commerce platforms. Although we may have accurate sentence representations, finding similar questions takes time. Our problem requires handling a high precision rate, as a misleading answer can be an inadequate experience for the customer.

A semantic search retrieval process finds questions already answered that have the same meaning as the new incoming question. This first stage acts as a filter to get possible similar candidates. Our solution converts sentences to vectors to enable a similarity computation and verifies that they are semantically identical; we perform this conversion with the Universal Sentence Encoder model large multilingual [Cer et al. 2018]. Then a semantic search occurs in a database using Elasticsearch [Gormley and Tong 2015]. This reduces the vector space. Finally, our trained neural network performs a classification task to evaluate pairs of questions’ similarity in these reduced spaces to obtain the sentence with the highest score (similar).

We deployed the system to online stores in the massive marketplace in Brazil and monitored the system for one month. We report results from May 14, 2023, to June 14, 2023. Our system helped to answer 13,991 input questions from real-world customers.

The evaluation results show that our system can answer questions that current chatbots have difficulties, increasing the total number of questions answered correctly, and it is suited to improving customers’ experience without needing manual labor.

The remainder of this article is organized as follows: Section 2 presents related studies. Section 3 reports on the full description of the developed system. Section 4 describes the experimental results in deploying and assessing the solution in the *GoBots* company environment, along with an automatic evaluation with annotated data. Section 5 discusses the obtained findings; Section 6 presents the conclusion and future work.

2. Related Work

Question-Answering systems are well-known Natural Language Processing applications; in literature [Kulkarni et al. 2019] propose a solution that analyses the product’s description and user’s review content by exploring semantic annotation based on ontologies, an intent classifier, and an answer ranking component.

[Chen et al. 2019] use sentence embedding with two multi-layer convolution networks: one to find relevant user review snippets and another to obtain the answer for the desired question from these snippets. [Gupta et al. 2019] that follows a similar with a review-based QA to synthesize the review and answer the customer questions. The solutions provided rely on user reviews and community answers, which may not always be entirely reliable due to their bias towards users’ opinions. The closest study to our proposal is the one built by [Sakata et al. 2019]. In their work, a suitable response is chosen based on the query-question similarity through a query and frequently asked question chart and their answers. Furthermore, they consider the query-answer relevance with a BERT [Devlin et al. 2019] based component. This solution produced relevant results.

However, the research focuses on a formal domain in contrast e-commerce context in marketplace platforms that handle several online stores. Besides, for the query-question component, a system focused on Japanese queries (*TSUBAKI* [Shinzato et al. 2008]) instead of universal ones was explored. A problem with this solution is that it requires an expensive GPU server to train and run, which can be a problem for smaller companies.

Moreover, [Mass et al. 2020] study FAQ retrieval using BERT to train two models to match *questions* and *answers* relevance from a given query. Furthermore, the authors proposed using question paraphrasing to overcome the size limitation of datasets.

[Gupta and Carvalho 2019] propose a multiple deep learning architecture based on attention [Vaswani et al. 2017] mechanism to compute both query-question and query-answer similarities to retrieve a response.

Alternatively, [Finardi et al. 2021] explore FAQ retrieval oriented to a specific domain constructing BERTaú, a Portuguese financial model, which learns a specific context representation. This model uses the data from the Itau Bank chatbots to train a neural network. The research applied the models for FAQ retrieval as a robust test over the objective domain, improving over other BERT models.

Although the existing approaches match question-answer relevance, e-commerce solutions might not benefit because historical stores' answers can be simple short sentences. For example, a simple answer might be “*yes it works*” and customized responses to specific questions such as “*Hi, thanks for buying with us. You can change the product if they have a problem*”. For actual e-commerce applications, this can derive from deploying specific solutions for each store, which is not scalable for big businesses. Thus, e-commerce FAQ retrieval likely finds more benefits from exploring only similarities across questions and criteria for deciding whether to answer.

3. Deep FAQ: A question-answering system based on sentence similarity

Our proposed solution called *DeepFAQ* is organized into two parts: *Buildtime* (cf. Subsection 3.1) and *Runtime* (cf. Subsection 3.2). The former consists of operations performed before the system answers a question; the latter consists of operations performed during the execution of the system to find an answer to a new incoming question. The proposed system works alongside another chatbot developed by *GoBots company*. From now on, we refer to this *company* system as the *Base System*. The *Base System* works by extracting intention and keyword entities from a given question and matching patterns between them with pre-made answers. The *DeepFAQ* system works on top of this solution, answering questions that the *Base System* cannot handle. Therefore, the *GoBots* team who manages the bots can control the question that goes to it or any other system parameter.

3.1. Buildtime

Before the system can reply to an incoming question for a given store in an e-commerce platform, it is necessary to generate a vector space from which it is possible to retrieve potential candidate questions. A neural network called *DeepSim*, which ranks questions on their similarity, is trained so that it is possible to use it on the *Runtime* (cf. Subsection 3.2). Figure 1 presents the components and procedures. The upcoming subsections explain each component in detail.

Question Answer retrieval and pre-processing: The first step involves building a vector space with potential candidate questions to retrieve the question-answer pairs from a store. To achieve this, we have created a script that accesses the *Mercado Livre API* and retrieves the questions from the store, obtaining all QA pairs in the store's history since the store joined the *Mercado Livre* e-commerce platform. Once we have the QA pairs from the store, we perform some basic pre-processing: removing stop words, converting the characters to lowercase, and removing accent marks. The process of removing stop words from answers is essential because customer service answers usually contain greetings such as daytime (*e.g.*, good morning, good night, *etc.*). Additionally, there exists a signature from the store's employee.

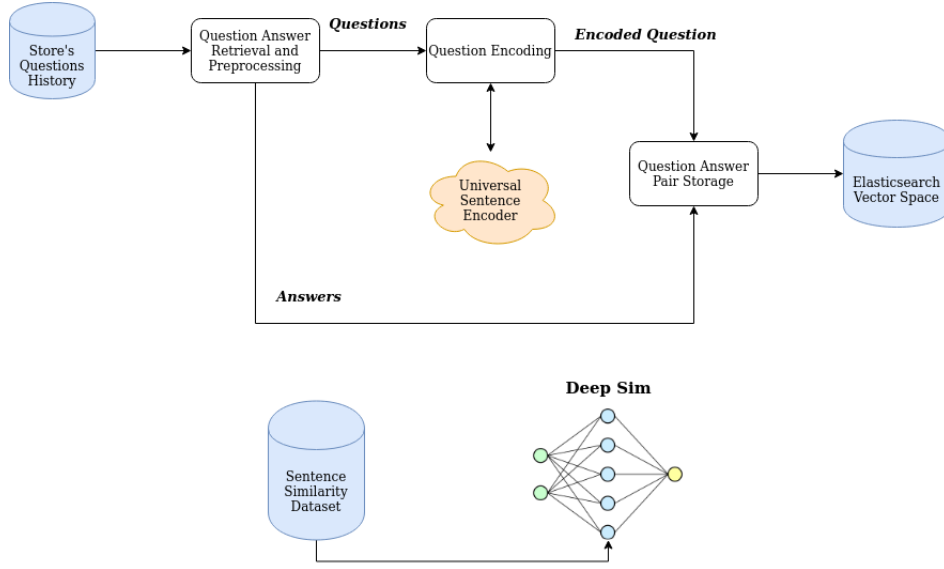


Figure 1. DeepFAQ's Buildtime procedure

Question Encoding: After cleaning QA, we encode these questions using the *Universal Sentence Encoder (USE)* proposed by [Cer et al. 2018]. Encoding involves sending questions to a remote service running to obtain their embedding and associate response. Our solution is suited to replace USE for any other encoders to be updated to improvements over the field of semantic representation.

Question-Answer Pair Storage: After obtaining the sentence embeddings, each QA pair and its question embeddings are sent to our vector space database with *Elasticsearch*. This avoids storing repeated questions so that the candidate retrieval step in *Runtime* is not affected (cf. Subsection 3.2). Finally, we store the QA pair and the question embedding, and the *product identifier* for which the asked question.

DeepSim Training: We trained a neural network called DeepSim alongside the QA pair database population process. We found the optimal architecture for a Siamese network with one layer, 512 neurons, relu activation, and 0.001 learning rate through random search, over this setup we trained using early stopping criteria of three over its loss function. This network compares two encoded questions and outputs the probability of them being identical, improving the system's accuracy for finding similar questions.

Finding Portuguese datasets for sentence similarity for e-commerce can be challenging; the few available, like ASSIN [Fonseca et al. 2016], are small and oriented for formal language, which is unsuitable for e-commerce queries. We decided to combine different kinds of datasets in this scenario. The first kind of data was the *Quora Question Pair* dataset ¹, we translated from English to Portuguese using the *AWS Translator* ².

The second kind of data was auto-generated pairs of sentences based on a pair of sentences template; those templates are common questions. We use a simple script to change selected words from the query to achieve similar or different questions. We ob-

¹<https://www.kaggle.com/c/quora-question-pairs>

²https://docs.aws.amazon.com/translate/latest/dg/API_TranslateText.html

tained a dataset from the DeepFAQ system by deploying a beta version to some stores and logging similar questions. We ended up with about 6000 pairs of questions annotated by a freelancer and verified by two others for reliability. All datasets contain two annotated sentences, whether they are similar or not.

3.2. Runtime

Figure 2 presents the system with the active components at the runtime of the solution. The process of answering a new incoming question explores the constructed vector space database (in build-time – cf. subsection 3.1) with the QAs history for a particular online store in the e-commerce platform. Our solution determines if two questions are similar or not. In the following sections, we explain how the system works in more detail.

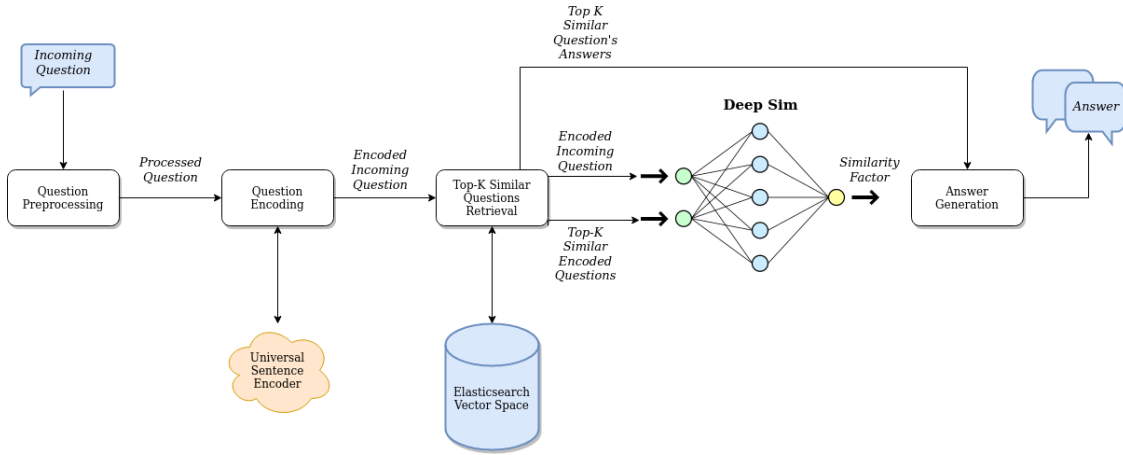


Figure 2. DeepFAQ's Runtime – deployed system ready to answer clients' new incoming NL questions in an online e-commerce store

Question pre-processing and encoding: The basic pre-processing performed on the incoming question is similar to the one made in the *Buildtime*. This processed question is then encoded with the *USE* model, obtaining the sentence embedding.

Top-K similar question retrieval: With the encoded question available, a query is performed in the database holding the vector space of previous questions; this procedure retrieves the top-k most similar questions using cosine similarity, and k is a parameter in the system. The set with the potential candidates holds their encoded questions and their QA pair in natural language. *Elasticsearch* does not implement natively an excellent way to search the top k questions, so we used the *AWS Elasticsearch Service* that is based on *Elasticsearch Open Distro* that comes with a KNN algorithm ready to use.

Ranking candidate questions via the *DeepSim* network: In some tasks, this first retrieval is enough to provide good results to the final user. However, in our case, we still need one more step further to answer questions correctly. The reason for this is that the vectors provided by [Yang et al. 2020], which is a multilingual version of *USE* encapsulates the intention of a question, but fail to notice that the object (ex. Product).

In our context, this is crucial because a particular question about a product *A* specification may have a completely different answer than a product *B*. Therefore, to solve this, each question from the set of top-k candidates is then used as input for the

DeepSim network in conjunction with the incoming encoded questions. This process enables the Deep-sim to rank each element in the top-k candidates by associating them with confidence, reflecting a possibility of being similar to the original incoming question.

Answer generation: Once the candidates are ranked, the solution selects the question candidate in the set with the *highest confidence to being similar*, given by *DeepSim*. The candidate’s question must satisfy a probability threshold set by the online store on the e-commerce platform to respond to the client’s question, which can change dynamically. Such question has an associate answer pair from the *Store’s Questions History* (Figure 1), we use this historical information to provide answers to our clients’ questions.

After generating an answer, the system performs post-processing by adding greetings and signatures from the store. Occasionally, responses may contain URLs, and the system checks if they are still available before providing the answer. Finally, if it has a problem, it chooses not to answer. After the system answers, a procedure called answer review is performed by human attendants from the stores to evaluate the answer and ask for its correction.

Continuous populating: When the system fails to answer a question, a human attendant answers the question on the platform as (s)he would do it if there were no automatic solution to this end. When the human does this, we implement a system that sends a notification to *DeepFAQ* and then stores this new question answered. This action allows us to be updated with new products, pairs of question and answers by promoting continuously the system evolution, and thus, answering more questions.

4. Experimental Evaluation

Our system was evaluated based on real-world stores that sell various products, such as clothes, car parts, electronics, and furniture. The evaluation covered the entire solution, currently deployed on the largest e-commerce platform in South America. All stores are utilizing our DeepFAQ tool in their production process. We gathered quantitative results to understand how crucial DeepFAQ is for these stores. Our main focus is to provide reports on the effectiveness of responses and the cash conversion rate connected to customer inquiries. Essentially, we analyzed whether questions answered by deepFAQ result in product sales.

4.1. Overall effectiveness on a real scenario

This analysis aimed to understand how many questions our system could answer and how these results automatize the human attendance process. In this sense, the objective of this analysis was to measure the impact of the proposed solution, observing how well it does in conjunction with the Base System. The collected data corresponded from stores of 338 clients. All stores have their vector space databases, which could have more than 3 million questions. For this experiment, we evaluated the efficiency of DeepFAQ in actual stores by gathering data for one month. We chose this timeframe because our solution’s application remained unchanged, with no significant software updates. In the period of our evaluation, DeepFAQ answered 13,991 questions.

Regarding its accuracy, since it would be very laborious to evaluate all these questions manually, we used the feedback from the stores themselves, which uses a review system to indicate if a response is correct or not. The revised question achieved a total

effectiveness of 97.75% from the questions answered by DeepFAQ during the evaluation time. Additionally, none of the stores using the system asked for it to be shut down in no time, indicating high user satisfaction.

Table 1 shows a summary of most common questions' intentions answered by DeepFAQ. These questions are related to specification, compatibility, availability, and others for intentions as greetings and acknowledge, which do not represent 1% of the total questions. We notice that the specification question means 41.33% of the question answered by DeepFAQ solutions. While compatibility, availability, and others categories represent 23.36%, 23.92%, and 11.39%, respectively. Moreover, DeepFAQ achieved a precision higher than 0.97 for each intention.

Table 1. Questions answered by DeepFAQ from May 14, 2023 to June 14, 2023

Client feedback		
Intention	n questions	Effectiveness (%)
Overall	13991	97.75
Specification	5783	97.75
Compatibility	3269	98.01
Availability	3346	97.46
Others	1593	97.80

4.2. Cash conversion using DeepFAQ by countries marketplaces

In this analysis, we utilized a recently implemented report to demonstrate to our clients (marketplaces) how our system can help to sell their products. We are reporting on the Cash Conversion between May 14, 2023, and June 14, 2023. The instability of the dollar exchange rate in the long term has made it challenging to measure long past periods, which is why we have chosen to focus on a shorter period.

Table 2 presents the cash conversion of the questions managed by DeepFAQ in the four countries where DeepFAQ operates; we use a money exchange provided by Google on June 20, 2023 to convert the local currency to USD. We notice a significant conversion in Brazil regions where DeepFAQ helped marketplaces obtain **\$45,177.07** while others regions got around 10% of this quantity; the total conversion amount is **\$58,683.17**.

Table 2. Cash conversion using DeepFAQ from May 14, 2023 to June 14, 2023

Cash conversion (\$USD)	
Argentina	4,939.07
Brazil	45,177.24
Chile	4,513.37
Mexico	4,053.49

5. Discussion

In our first analysis, we observed that *DeepFAQ* can perform well in a real-world scenario over a one-month period. Even though the number of answer question answered by DeepFAQ is "low" (13991), this can be explained by the absence of appropriate candidates for

a given question. A DeepFAQ solution relies on the existing appropriate candidates in the space to provide relevant answers. However, our solutions fix such problems for future questions applying the *Continuous populating* described in Section 3.

Our experiments showed that the system’s answer effectiveness is **97.75%**, which reveals that our system can usually give a correct answer. Additionally, Table 1 indicated that client satisfaction remains high across questions with different intentions, demonstrating that our solution is not biased toward specific questions.

Marketplaces are interested in understanding how our solutions can assist them in resolving customer inquiries. Also, they want to know if our solutions generate sales for their stores, measuring their monetary value. Our solutions provide adequate answers to multiple intentions of questions, which makes them suitable for retrieving FAQs from the e-commerce domain. Our quality analysis indicated a high keyword sensibility that only allows answering a question with the most ranked candidate.

Our analysis demonstrated that the presence of DeepFAQ is helpful during the sales, being a crucial key factor because its absence can have noticeable effects, especially for Brazilian marketplaces in a few months. The low conversion rate in countries outside of Brazil may be attributed to the fact that the *GoBots* company initially began with Brazilian stores, which make up approximately 80% of all DeepFAQ-affiliated stores.

6. Conclusion

Question answering based on sentence similarity computation remains an open research challenge. This requires solutions to handle several domains simultaneously and adequately addresses informal issues raised in Portuguese language questions. This study presented and evaluated a system by combining semantic search with a neural network as a classification system to rank similarity scores. Our system retrieves similar questions based on vector space encoding. The proposal showed an effective solution to help to solve the problem of question-answering in e-commerce platforms with a small need for data labeling or large languages models services. Furthermore, working with existing software systems implemented in this context made it possible to significantly increase the number of questions answered by maintaining a good enough precision, considering the 97.75% of assertiveness in the answers given. Our evaluation process in a real-world setting showed that our solution is an up-and-coming technique to contribute to addressing question-answering in e-commerce platforms. In future work, we plan to overcome the problems of auto-populating Questions-Answers pair using Large Language Models. This might reduce the need to have human historical questions-answer pair and anticipate the need for inquiries from customers.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. In addition, this research was partially funded by the São Paulo Research Foundation (FAPESP) (grant #2022/13694-0).³ We would like to thank GoBots for collecting data and sharing their environment.

³The opinions expressed in this work do not necessarily reflect those of the funding agencies.

References

- Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strophe, B., and Kurzweil, R. (2018). Universal sentence encoder.
- Chen, S., Li, C., Ji, F., Zhou, W., and Chen, H. (2019). Review-driven answer generation for product-related questions in e-commerce. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19*, page 411–419, New York, NY, USA. Association for Computing Machinery.
- 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.
- Finardi, P., Viegas, J. D., Ferreira, G. T., Mansano, A. F., and Carid’a, V. F. (2021). Bertaú: Itaú bert for digital customer service. *ArXiv*, abs/2101.12015.
- Fonseca, E. R., Borges dos Santos, L., Criscuolo, M., and Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.
- Gormley, C. and Tong, Z. (2015). *Elasticsearch: The Definitive Guide*. O’Reilly Media, Inc., 1st edition.
- Gupta, M., Kulkarni, N., Chanda, R., Rayasam, A., and Lipton, Z. C. (2019). Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.
- Gupta, S. and Carvalho, V. R. (2019). Faq retrieval using attentive matching. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’19*, page 929–932, New York, NY, USA. Association for Computing Machinery.
- Kulkarni, A., Mehta, K., Garg, S., Bansal, V., Rasiwasia, N., and Sengamedu, S. (2019). Productqna: Answering user questions on e-commerce product pages. In *Companion Proceedings of The 2019 World Wide Web Conference, WWW ’19*, page 354–360, New York, NY, USA. Association for Computing Machinery.
- Mass, Y., Carmeli, B., Roitman, H., and Konopnicki, D. (2020). Unsupervised FAQ retrieval with question generation and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812, Online. Association for Computational Linguistics.
- Sakata, W., Shibata, T., Tanaka, R., and Kurohashi, S. (2019). Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’19*, page 1113–1116, New York, NY, USA. Association for Computing Machinery.
- Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C., and Kurohashi, S. (2008). TSUB-AKI: An open search engine infrastructure for developing new information access

methodology. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

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.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strophe, B., and Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Classificação de Polaridade Orientada aos Alvos de Opinião em Comentários sobre Debate Político em Português

Eloize R. Marques Seno¹, Fábio S. Igarashi Anno¹,
Lucas Lazarini¹, Helena M. Caseli²

¹Instituto Federal de São Paulo – São Carlos, SP

²Departamento de Ciência da Computação – Universidade Federal de São Carlos
São Carlos, SP

eloize@ifsp.edu.br, {fabio.seyiji, lazarini.lucas}@aluno.ifsp.edu.br,

helenacaseli@ufscar.br

Abstract. *The internet and, especially social media, are fertile ground for publishing opinions on various subjects, products, and services. Traditionally, automatic analysis of opinions has been based on words that denote polarity or emotion. However, with the emergence of large language models like ChatGPT, the way in which we process text for subjective analysis has changed a lot. In this context, this paper aims to investigate the potential of ChatGPT – compared to a neural model for emotion identification in texts, and lexicon-based approaches – in polarity classification oriented towards opinion targets in comments on political debate in Portuguese.*

Resumo. *A internet e, em especial as mídias sociais, são um terreno fértil para a publicação de opiniões sobre os mais diversos assuntos, produtos e serviços. Tradicionalmente a análise automática de opiniões é realizada com base nas palavras que denotam alguma polaridade ou emoção. Contudo, com o surgimento dos grandes modelos de linguagem, como o ChatGPT, a maneira com a qual processamos textos para realizar análises subjetivas mudou bastante. Neste contexto, este artigo tem como foco investigar as potencialidades do ChatGPT – comparada com a de um modelo neural de identificação de emoções em textos e de abordagens baseadas em léxico – na classificação de polaridade orientada para os alvos de opinião em comentários sobre debate político em português.*

1. Introdução

O enorme volume de textos opinativos compartilhados nas mídias sociais como textos de avaliação de produtos em sites de comércio eletrônico, comentários de notícias sobre política, publicações em redes sociais, etc. tem despertado um interesse cada vez maior por algoritmos capazes de analisar a opinião pública. Essas opiniões e sentimentos têm grande apelo comercial e acadêmico, por influenciar nas tomadas de decisões dos indivíduos.

A análise de opiniões expressas em textos é o objeto de estudo da área de Análise de Sentimentos (AS), ou Mineração de Opiniões, uma subárea do Processamento de Linguagem Natural (PLN) que visa a utilização de recursos e ferramentas linguístico-computacionais para analisar automaticamente as opiniões e sentimentos de

peessoas com relação a entidades como produtos, serviços, organizações e indivíduos [Liu and Zhang 2012]. A AS se divide em duas tarefas principais: (i) mineração de opinião e (ii) mineração de emoção. A mineração de opinião está relacionada à detecção de polaridade (ou valência) associada à opinião do autor de uma postagem a respeito de uma entidade ou assunto em discussão, podendo ser positiva, negativa ou neutra. A mineração de emoção refere-se à identificação da emoção associada ao texto da postagem produzido pelo autor, como alegria, surpresa, raiva ou tristeza [Yadollahi et al. 2017]. Este estudo, mais especificamente, tem como foco a mineração de opinião (daqui em diante, classificação de polaridade) em textos de comentários do domínio político.

Tradicionalmente, a classificação de polaridade se refere ao sentimento geral conduzido por um documento ou sentença em particular. Contudo, as abordagens que analisam o sentimento geral de um documento/sentença podem apresentar limitações, quando um mesmo texto expressa opiniões sobre diferentes entidades, chamadas de alvos de opinião. Por exemplo, em um dos comentários presente no corpus utilizado neste estudo: “Portas muito bem, Sócrates a patinar.”, o sentimento em relação ao alvo “Portas” é positivo, enquanto o sentimento em relação ao alvo “Sócrates” é negativo. Para tratar adequadamente esses casos, é necessário considerar os diferentes alvos de opinião expressos em um mesmo comentário. É neste contexto que este trabalho se insere. Mais especificamente, o foco de interesse é a classificação de polaridade orientada para os alvos de opinião em textos políticos escritos em português. Ou seja, uma tarefa mais específica e desafiadora do que a análise geral do sentimento expresso pelo texto todo e menos refinada que a análise de sentimentos baseada em aspectos (por exemplo, [Assi et al. 2022, Schouten and Frasnica 2016]).

Para a classificação de polaridade orientada aos alvos de opinião em comentários sobre política, neste trabalho foram investigadas as potencialidades e limitações de uso do ChatGPT e de um modelo BERT de identificação de emoções, conhecido por GoEmotions, pré-treinado para o português [Hammes and Freitas 2021]. Além desses modelos, abordagens baseadas em léxicos de sentimentos – LIWC [Balage Filho et al. 2013] e SentiLex-PT [Carvalho and Silva 2015] – foram usadas como base de comparação. Os resultados de um experimento preliminar mostraram que o ChatGPT tem um potencial maior na realização da tarefa proposta, comparado aos demais modelos, apesar de não ser muito estável nas respostas. Até onde se sabe, este é o primeiro trabalho para o português a investigar o uso do ChatGPT na classificação de sentimento orientado ao alvo de opinião em textos em português.

O restante deste artigo está organizado da seguinte forma. A seção 2 apresenta de forma breve alguns trabalhos da literatura relacionados a este. Na seção 3 são apresentadas as estratégias investigadas para a classificação de sentimentos orientada aos alvos de opinião em textos, bem como o corpus de trabalho e os recursos e ferramentas linguístico-computacionais usados. Os resultados obtidos com cada estratégia são sumarizados na seção 4. Por fim, a seção 5 apresenta as principais conclusões deste trabalho.

2. Trabalhos relacionados

Na literatura, os trabalhos relacionados à classificação de polaridade tem como foco a análise do sentimento geral de documentos ou sentenças (por exemplo, vide [França and Oliveira 2014], [Araújo et al. 2020] e [Capellaro and Caseli 2021]) ou de um

aspecto (característica) específico de uma entidade (por exemplo, [Assi et al. 2022] e [Schouten and Frasincar 2016]).

Muitas abordagens distintas já foram investigadas [Liu and Zhang 2012, Pereira 2021, Hung and Alias 2023]. Dentre elas destacam-se as seguintes abordagens: (i) baseadas em conhecimento, que identificam palavras e expressões que carregam sentimentos a partir de léxicos construídos com base em *corpus* ou em recursos como o Wordnet (por exemplo, [Taboada et al. 2011] e [Akilandeswari and Jothi 2018]), (ii) baseadas em modelos tradicionais de aprendizado de máquina supervisionados ou não supervisionados (por exemplo, [Carvalho et al. 2017] e [Araújo et al. 2020]) e (iii) baseadas em aprendizado profundo [Zhang et al. 2018]. Abordagens híbridas combinando essas técnicas também são comuns [Appel et al. 2016].

Em [Carvalho et al. 2017], por exemplo, os autores compararam o desempenho de três algoritmos de aprendizado supervisionado (Naive Bayes, SVM e MaxEnt) e de três métodos de seleção de atributos (Qui Quadrado, CPD e CPPD) na classificação de polaridade em parágrafos de textos relacionados às eleições de 2014 no Brasil. A melhor performance, medida em termos de acurácia (cerca de 85%), foi obtida com o classificador MaxEnt usando o CPPD como método de seleção de atributos. Já em [Capellaro and Caseli 2021] os autores usaram modelos do tipo BERT, baseados em aprendizado profundo, na análise de polaridade e de tópicos em *tweets* relacionados às eleições de 2018 no Brasil. Usando um modelo pré-treinado para o português, a abordagem dos autores alcançou um *F1-score* de 96,6% na classificação da polaridade associada ao sentimento geral dos *tweets*.

Um trabalho que tem maior proximidade com esta pesquisa é o de [Assi et al. 2022]. Usando o modelo GoEmotions para o português [Hammes and Freitas 2021], os autores analisaram as emoções relacionadas a aspectos específicos de entidades em textos de revisão do setor de acomodação e, posteriormente, mapearam cada emoção para uma das três polaridades possíveis (positiva, negativa ou neutra), obtendo, assim, a classificação de polaridade para cada aspecto de uma entidade particular presente no texto. A estratégia dos autores alcançou um *F1-score* de 61,0%. De maneira similar a realizada por [Assi et al. 2022], neste trabalho o modelo pré-treinado por [Hammes and Freitas 2021] foi também usado para a identificação de emoções em comentários sobre debate político e, posteriormente, as emoções detectadas foram mapeadas para polaridades (conforme Seção 3). Porém, diferentemente de [Assi et al. 2022] nesta pesquisa a AS é realizada em nível de sentenças e não de aspectos.

3. Análise de sentimentos orientada aos alvos de opinião

Para a classificação de sentimentos (polaridade) orientada aos alvos de opinião em comentários sobre debate político em português, foi investigado o uso do ChatGPT e do modelo GoEmotions [Hammes and Freitas 2021]. A versão usada do ChatGPT se refere ao modelo GPT 3.5, disponível *online*¹ no período de 24 a 28 de junho de 2023. O GoEmotions, por sua vez, é um ajuste fino do BERTimbau [Souza et al. 2020] para a classificação de emoções em português. Esse modelo produz como saída uma lista de emoções ranqueadas de acordo com um valor probabilístico. São 27 emoções possíveis

¹Disponível em: <https://chat.openai.com/>, versão de 24 de Maio, segundo a página. (Acessado em: 28/06/2023).

mais a emoção “neutra”. Quanto maior o valor associado à uma emoção, maior a probabilidade dela ocorrer no texto. Assim, para associar uma polaridade aos alvos de opinião, cada emoção foi mapeada para um sentimento positivo (1), negativo (-1) ou neutro (0), conforme apresentado na Tabela 1.

Tabela 1. Mapeamento das emoções retornadas pelo GoEmotions para polaridades

Polaridade	Emoções
-1 (negativo)	aborrecimento, constrangimento, decepção, desaprovação, luto, medo, nervosismo, nojo, raiva, remorso e tristeza
0 (neutro)	neutro, curiosidade, confusão, percepção e surpresa
1 (positivo)	admiração, alegria, alívio, aprovação, amor, desejo, diversão, entusiasmo, gratidão, orgulho, otimismo e zelo

Além dos modelos GPT e GoEmotions, também foram implementadas estratégias baseadas em léxicos de sentimentos, que serviram de base de comparação com os dois modelos neurais. Mais especificamente, foram usados os léxicos LIWC-PT² [Balage Filho et al. 2013] e SentiLex-PT³ [Carvalho and Silva 2015]. O LIWC-PT é um léxico geral do português constituído por 127.149 instâncias organizadas em categorias. As categorias *posemo* e *negemo* indicam emoção (polaridade) positiva e negativa, respectivamente. Há também categorias para mineração de emoções como afeto, raiva, tristeza, etc., porém essas não foram exploradas neste estudo. O SentiLex-PT, por sua vez, foi concebido especificamente para a análise de sentimento e opinião sobre entidades humanas. Ele é composto por 7.014 lemas e 82.347 formas flexionadas organizadas em adjetivos, substantivos, verbos e expressões idiomáticas. Nos experimentos apresentados neste artigo, as expressões idiomáticas não foram usadas.

A subseção 3.1 apresenta uma descrição das abordagens investigadas neste trabalho e a subseção 3.2 descreve o corpus usado na avaliação de cada abordagem.

3.1. Descrição das abordagens de classificação de polaridade

A primeira abordagem é baseada no modelo GPT. Em linhas gerais, o ChatGPT recebe um texto de entrada (daqui em diante, *prompt*) com a descrição da tarefa a ser realizada pelo sistema e gera as saídas conforme solicitado. O desafio principal em lidar com o ChatGPT consiste em definir um *prompt* que gere as saídas exatamente como se espera para determinada tarefa. Essa “engenharia de *prompt*” não se mostrou uma tarefa simples, pois o sistema é bastante variável e por vezes até contraditório. Depois de algumas tentativas usando uma pequena amostra do corpus (10 sentenças), chegou-se ao seguinte *prompt*: “Dada a sentença <sentença> e o alvo sendo <alvo> qual a polaridade associada a ele, negativa (-1), positiva (1) ou neutra (0)?”, sendo <sentença> a sentença do comentário e <alvo> o alvo de opinião marcado no corpus (conforme explicado na Seção 3.2).

²Disponível em: <http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc>. Acesso em: 28 jun. 2023.

³Disponível em: <https://github.com/sillasgonzaga/lexiconPT>. Acesso em: 27 jun. 2023.

A segunda abordagem é baseada no GoEmotions. Para este modelo duas estratégias também foram investigadas: (BASE) a que considera o comentário como um todo para predição da emoção e (VIZ) a que considera apenas a região de vizinhança seguindo abordagem similar à de [Assi et al. 2022]. A polaridade atribuída é aquela resultante da combinação das top-3 melhores emoções retornadas com pontuação maior ou igual a 0,3, mesmo limiar adotado em [Assi et al. 2022].

Por fim, a terceira abordagem investigada faz uso de léxicos (LIWC-PT ou SentiLex-PT). Essa abordagem consistiu em atribuir ao alvo de opinião do comentário a polaridade resultante da soma das polaridades das palavras presentes no léxico. Neste sentido, duas estratégias de atribuição de polaridade foram investigadas: (BASE) uma que considera o comentário como um todo e (DEP) uma que considera apenas as palavras associadas ao alvo do comentário, já fornecido para cada sentença do corpus (Seção 3.2), via dependência sintática.⁴ Assim, para cada palavra com polaridade presente no léxico (SentiLex-PT ou LIWC-PT) que também estava presente no comentário recuperou-se sua polaridade. Na estratégia BASE considerou-se como resultado a soma das polaridades de todas as palavras e no caso da estratégia DEP apenas as polaridades das palavras associadas ao alvo por alguma relação de dependência sintática.

3.2. Corpus de investigação

Como corpus de trabalho foi usado o SentiCorpus-PT [Carvalho et al. 2011]. A escolha desse corpus se deu pelo fato de que ele dispõe de anotações de polaridade de referência (isto é, realizada por humanos) e orientada aos alvos de opiniões presentes nos textos. O SentiCorpus-PT é composto por comentários sobre debates televisivos referentes às eleições de 2009 do Parlamento Português. A versão usada do corpus é composta de 1.082 comentários, totalizando 3.868 sentenças anotadas. Porém, neste estudo foram usadas apenas as primeiras 200 sentenças do corpus. Este número limitado se deve ao fato de que a versão livre do ChatGPT não dispõe de API e, portanto, o processamento do corpus é feito manualmente, ou seja, os comentários são fornecidos em pequenas quantidades como entrada para o sistema.

Cada sentença no corpus pode ter diferentes alvos de opinião. Os alvos de opinião são entidades humanas, nomeadamente políticos, organizações políticas (geralmente utilizadas para se referir aos seus membros), personalidades da mídia (por exemplo, jornalistas) ou usuários (comentadores). Cada sentença do corpus dispõe de anotações de polaridade variando de -2 (o valor negativo mais forte) até 2 (o valor positivo mais forte) e de alvos de opinião. Entretanto, nos experimentos realizados neste estudo a polaridade -2 foi mapeada para -1 (negativa) e a polaridade 2 foi mapeada para 1 (positiva).

4. Avaliação das estratégias

A avaliação das estratégias investigadas nesta pesquisa foi realizada tendo como referência a anotação manual de polaridades associadas aos alvos, fornecida pelo SentiCorpus-PT. Apesar do corpus dispor da anotação de polaridade orientada aos alvos de opiniões, optou-se por fazer uma nova anotação, a fim de verificar a concordância entre a anotação original feita por falantes do português europeu e a anotação realizada

⁴A dependência sintática neste caso foi gerada pelo modelo `pt_core_news_lg` do Spacy, disponível em: <https://spacy.io/models/pt>. Acesso em: 30 jun. 2023.

por falantes nativos do português na variante falada no Brasil. Os anotadores foram dois autores deste trabalho com larga experiência em anotação de corpus. A Figura 1 traz a distribuição de polaridade atribuída pela Referência (anotação original), pelos humanos falantes do português do Brasil e por cada uma das estratégias investigadas neste trabalho.



Figura 1. Distribuição de classes da Referência (anotação original), dos brasileiros e das estratégias investigadas neste trabalho

Como é possível notar na Figura 1, a anotação de Referência tem a maioria dos alvos associados à polaridade negativa. Os dois anotadores humanos (brasileiros) e o ChatGPT geraram uma distribuição de classes similar à da Referência, porém com uma quantidade maior de neutros do que ela. Acredita-se que o fato da anotação original ter sido realizada por anotadores que tinham um conhecimento do contexto do debate realizado em Portugal pode ter influenciado a atribuição de polaridades negativas mesmo quando o comentário em si não trazia explicitamente palavras ou sentimentos que denotassem essa polaridade.

O gráfico também traz a distribuição de classes para as abordagens baseadas em léxicos (SL e LW) e emoções (GE). Neste caso, nota-se que o SentiLex é mais balanceado, o LIWC tende para o positivo e o GoEmotions, para o neutro. As estratégias que limitam a associação de palavras de polaridade àquelas ligadas por alguma relação de dependência sintática ao alvo (DEP) levaram a um número muito grande de alvos não associados a polaridades (N/A); enquanto que a análise de emoção usando apenas a região de vizinhança do alvo (GE-VIZ) diminuiu a quantidade de alvos associados à polaridade negativa e aumentou a quantidade de alvos associados à polaridade neutra. Por fim, na parte inferior da figura são apresentados os valores para as combinações das estratégias, sendo que combinou-se as estratégias base (BASE) entre elas e as estratégias que consideram o alvo na determinação da polaridade, seja por dependência sintática (DEP) ou região de vizinhança (VIZ), entre elas.

A Tabela 2 traz os resultados da comparação das saídas geradas por cada estratégia

investigada neste artigo em comparação com a referência. Os resultados (porcentagens) são apresentados para cada classe (positiva, negativa e neutra) em termos de Precisão, Cobertura e F1-score. Seja c uma classe qualquer, a Precisão considera a quantidade de instâncias da classe c corretamente classificadas em relação ao total de instâncias classificadas como sendo da classe c . A Cobertura representa a quantidade de instâncias da classe c corretamente classificadas sobre o total de instâncias classificadas como classe c na anotação de referência. Por fim, F1-score representa a média harmônica entre a Precisão e a Cobertura. Após obter o F1-score de cada classe, calculamos a média aritmética entre os valores de F1-score das três classes, representada pela medida Macro-F1.

Tabela 2. Resultados das estratégias investigadas neste trabalho em termos de Precisão (P), Cobertura (C), F1-score e Macro-F1 (M-F1)

	Positiva			Negativa			Neutra			
	P	C	F1	P	C	F1	P	C	F1	M-F1
Humano 1	85,0	85,0	85,0	100,0	100,0	100,0	83,8	44,3	57,9	82,1
Humano 2	84,2	80,0	85,0	82,0	98,0	73,0	83,6	43,3	96,3	75,1
ChatGPT	72,7	80,0	76,2	88,3	51,1	64,8	23,4	66,7	34,6	58,5
SL-BASE	35,7	37,5	36,6	76,8	32,3	45,5	10,0	14,8	11,9	31,3
LW-BASE	26,0	60,0	36,4	64,1	18,8	29,0	5,3	3,7	4,3	23,3
GE-BASE	53,8	35,0	42,4	81,8	13,5	23,2	18,6	96,3	31,1	32,3
SL+LW-BASE	25,6	55,0	34,9	73,7	55,0	63,0	6,0	55,0	10,9	36,3
SL+GE-BASE	38,0	47,5	42,2	79,6	35,3	49,0	18,0	59,3	27,6	39,6
LW+GE-BASE	27,0	60,0	37,2	68,9	23,3	34,8	9,4	60,0	16,2	29,4
SL+LW+GE-BASE	27,6	60,0	37,8	73,3	33,1	45,6	7,7	60,0	13,6	32,3
SL-DEP	66,7	5,0	9,3	84,6	8,3	15,1	0,0	0,0	0,0	12,2
LW-DEP	60,0	15,0	24,0	85,7	4,5	8,6	85,7	0,0	0,0	10,9
GE-VIZ	68,4	32,5	44,0	83,3	7,5	13,8	16,2	92,6	27,6	28,5
SL+LW-DEP	72,7	20,0	31,4	86,7	9,8	17,6	50,0	3,7	6,9	18,6
SL-DEP+GE-VIZ	66,7	35,0	45,9	83,3	15,0	25,5	16,2	88,9	27,4	32,9
LW-DEP+GE-VIZ	60,9	35,0	44,5	88,9	12,0	21,2	16,0	88,9	27,1	30,9
SL+LW-DEP+GE-VIZ	65,2	37,5	47,6	88,0	16,5	27,8	16,4	88,9	27,7	34,4

Como é possível notar na Tabela 2, o ChatGPT apresentou o melhor desempenho global (i.e. 58,5% de Macro-F1), comparado a todas as estratégias avaliadas, ou seja, o modelo GPT foi o que mais se aproximou dos valores obtidos pelos humanos. O segundo melhor desempenho (ou seja, 39,6%) foi alcançado com a combinação do SentiLex-PT e o GoEmotions sem usar as dependências sintáticas (SL+GE-BASE). Os piores desempenhos foram obtidos pelas estratégias baseadas no SentiLex e LIWC usando as dependências sintáticas (i.e. SL-DEP, LW-DEP e SL+LW-DEP). Entretanto, a estratégia SL+LW-DEP, que combina os dois léxicos com as dependências sintáticas, alcançou um desempenho muito similar ao do ChatGPT em termos de Precisão na classificação de sentimentos positivos e negativos e obteve uma Precisão ainda superior à daquele modelo na classificação de sentimentos neutros. Porém, essa estratégia obteve um desempenho muito ruim em termos de Cobertura, o que prejudicou o seu desempenho global. A baixa cobertura apresentada pelas estratégias que usaram dependência sintática pode estar relacionada ao fato de que vários comentários no corpus não apresentam a estrutura

esperada de sujeito-verbo-objeto, o que pode ter impactado na qualidade da análise de dependência gerada, fazendo com que a palavra com polaridade não fosse corretamente associada ao alvo do comentário. Quando se observa o desempenho de cada estratégia em relação a cada classe de forma isolada, percebe-se que os melhores valores de *F1-score* para todas as classes foram alcançados com o modelo GPT, sendo que ele teve um desempenho melhor na classificação de sentimentos positivos do que negativos e neutros. De maneira geral, todas as estratégias tiveram um pior desempenho na classificação de sentimentos neutros, sendo que a estratégia SL-DEP, baseada no SentiLex e nas dependências sintáticas, não conseguiu classificar corretamente nenhuma das instâncias da classe neutra. Para esta estratégia, especificamente, o valor de Macro-F1 foi calculado considerando os *F1-scores* apenas das classes positiva e neutra.

Tabela 3. Exemplo de comentário e as saídas retornadas pelas estratégias investigadas neste trabalho (IND representa as palavras de sentimento)

Comentário: Foi de facto um debate cordato, civilizado, em que Jerónimo se mostrou um senhor e o Louçã meteu a viola no saco.

Alvo: Jerónimo						Polaridade: 1
SL-BASE 1	SL-IND cordato, civilizado	LW-BASE -1	LW-IND debate, viola	GE-BASE 1	GE-EMO neutro, admiração	CHATGPT 1
SL-DEP N/A	SL-IND –	LW-DEP N/A	LW-IND –	GE-VIZ 0	GE-EMO neutro	

Alvo: Louçã

Alvo: Louçã						Polaridade: -1
SL-BASE 1	SL-IND cordato, civilizado	LW-BASE -1	LW-IND debate, viola	GE-BASE 1	GE-EMO neutro, admiração	CHATGPT -1
SL-DEP N/A	SL-IND –	LW-DEP N/A	LW-IND –	GE-VIZ 0	GE-EMO neutro	

A título de ilustração a Tabela 3 traz um exemplo de comentário do corpus e as saídas retornadas pelas estratégias investigadas neste trabalho.

5. Conclusão

Este estudo avaliou as potencialidades do ChatGPT na tarefa de análise de polaridade orientada para os alvos de opinião em comentários do domínio político, comparando seu desempenho com o de outro modelo neural de classificação de emoções em textos e com o de estratégias baseadas em léxicos. Experimentos preliminares usando um conjunto de dados pequeno (200 comentários) mostraram que o ChatGPT tem potencial para identificar a polaridade associada a cada alvo de opinião em um comentário, sendo esta a estratégia que mais se aproxima do desempenho obtido por humanos na mesma tarefa. Contudo, cabe ressaltar que o uso do ChatGPT apresenta vários desafios como (i) a escolha do *prompt* de entrada apropriado com a descrição da tarefa a ser realizada pelo sistema, crucial para que ele entenda o que o usuário espera como saída, (ii) a variabilidade das respostas dadas para uma mesma entrada para usuários distintos e (iii) a necessidade de relizar o processo manualmente na versão gratuita.

Como trabalho futuro pretende-se avaliar um conjunto de dados maior e experimentar outros analisadores sintáticos de dependência, uma vez que os maiores valores de precisão foram alcançados pelas estratégias que usaram dependência sintática para atribuir o sentimento ao alvo. Porém, elas apresentaram uma cobertura muito baixa.

Referências

- Akilandeswari, J. and Jothi, G. (2018). Sentiment classification of tweets with non-language features. *Procedia Computer Science*, 143:426–433. 8th International Conference on Advances in Computing Communications (ICACC-2018).
- Appel, O., Chiclana, F., Carter, J., and Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124. *New Avenues in Knowledge Bases for Natural Language Processing*.
- Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102.
- 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., 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.
- Capellaro, L. and Caseli, H. M. (2021). Análise de polaridade e de tópicos em tweets no domínio da política no Brasil. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 47–55, Porto Alegre, RS, Brasil. SBC.
- Carvalho, C. M. A., Nagano, H., and Barros, A. K. (2017). A comparative study for sentiment analysis on election Brazilian news. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 103–111, Uberlândia, Brazil. Sociedade Brasileira de Computação.
- Carvalho, P., Sarmiento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 564–568, Portland, Oregon, USA. Association for Computational Linguistics.
- Carvalho, P. and Silva, M. (2015). SentiLex-PT: Principais características e potencialidades. *Linguística, Informática e Tradução: Mundos que se Cruzam, Oslo Studies in Language*, 7(1):425–438.
- França, T. and Oliveira, J. (2014). Análise de sentimento de tweets relacionados aos protestos que ocorreram no Brasil entre junho e agosto de 2013. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 128–139, Porto Alegre, RS, Brasil. SBC.
- 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.

- Hung, L. and Alias, S. (2023). Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 27:84–95.
- Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *Mining Text Data*, pages 415–463.
- Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.
- Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.
- 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.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.
- Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, 50(2):1–33.
- Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8.

How Good Is ChatGPT For Detecting Hate Speech In Portuguese?

Amanda S. Oliveira¹, Thiago C. Cecote¹, Pedro H. L. Silva¹, Jadson C. Gertrudes¹,
Vander L. S. Freitas¹, Eduardo J. S. Luz¹

¹Computing Department – Federal University of Ouro Preto

{amanda.oliveira2, thiago.cecote}@aluno.ufop.edu.br

{silvap, jadson.castro, vander.freitas, eduluz}@ufop.edu.br

Abstract. *This study evaluates OpenAI’s ChatGPT, a large language model, for its efficacy in detecting hate speech in Portuguese tweets, comparing it with purpose-trained models. Despite incurring considerable computational costs, ChatGPT as a zero-shot classifier demonstrated commendable performance, even superior to or on par with state-of-the-art methods, with an F1-score of 73.0% on the ToLD-BR. In a cross-dataset evaluation on the HLPHP dataset, it secured a superior F1-score of 73%. The choice of prompt significantly impacts the outcome, with a wider scope prompt balancing precision and recall metrics. ChatGPT, due to its interpretability and resilience against data distribution shifts, could be a preferred choice for tasks prioritizing these factors.*

1. Introduction

The considerable increase in users and time spent on social media has led to a growing prevalence of hate attacks, as the false sense of anonymity encourages individuals to post derogatory comments online. Hate speech encompasses messages that convey intolerance or aversion towards specific groups, such as ethnic, religious, sexual, or gender minorities and immigrants. It may include derogatory comments, threats, and incitements to violence [Walther 2022].

Given the importance of detecting hate speech on social media in recent years, particularly as more toxic posts on networks like Twitter tend to receive higher engagement [Salehabadi et al. 2022], it is crucial to implement content moderation rules to prevent the sharing of toxic content.

Hate speech detection on social media can be viewed as a text classification problem involving natural language processing (NLP) techniques and machine learning to identify and filter offensive messages. Researchers have addressed this issue in various languages, including Arabic [Mubarak et al. 2017], German [Wiegand et al. 2018], and English [Davidson et al. 2017, Zampieri et al. 2019, Mandl et al. 2019], highlighting the need to consider each language’s structural and social aspects [Radfar et al. 2020]. This work focuses on the Portuguese language.

Advancements in Portuguese hate speech detection have been marked, with several studies contributing significantly. [de Pelle and Moreira 2017] compiled a corpus of 1,250 comments, mostly offensive, from a news site and established a benchmark using naive Bayes and SVM classifiers, the latter outperforming with a 77-82 F1-score. A European-Portuguese dataset of 5,668 tweets was developed by [Fortuna et al. 2019], labeled by

annotators of varying expertise using binary and hierarchical schemes. The dataset’s utility was demonstrated using pre-trained embeddings and LSTM.

Further advancements include the categorization of 7,000 Instagram comments as hate speech/offensive or non-offensive by [Vargas et al. 2022], using n-grams and bag-of-n-grams with tf-idf preprocessing and achieving an F1-score of 85% for hate speech detection and 78% for offensive speech. [Leite et al. 2020] introduced a large-scale Brazilian Portuguese dataset, ToLD-Br, containing 21,000 annotated tweets. BERT-based models were applied, achieving a macro-F1 score of 76% in binary mode. Despite these developments, there is room for improvement of large-scale monolingual data.

With the growing public interest in generative pre-trained transformer (GPT) models, such as OpenAI’s ChatGPT [Brown et al. 2020] ¹, it is natural for this type of model to be employed for various natural language tasks, including toxicity analysis in social media texts. However, Large Language Models (LLM) like ChatGPT are designed to be generalists and function as a chatbot. In this work, we investigate three research questions: (i) As a chatbot model, can ChatGPT effectively detect hate/toxic speech in social media texts? (ii) How well does ChatGPT perform compared to models specifically trained for hate/toxicity detection tasks? And (iii) what is the impact of the prompt? To address these questions, we propose a methodology for classifying tweets as hate speech or non-hate using the ChatGPT API in a zero-shot classification fashion. Additionally, we compare ChatGPT’s performance against a published baseline and a naive approach.

Our findings suggest the promising feasibility of employing ChatGPT to classify toxic/hateful textual content within Portuguese tweets. Furthermore, ChatGPT provides a more interpretable solution, a characteristic of considerable desirability, albeit at a significantly elevated computational cost. Hence, it may not represent the most optimal solution, given that alternative methods, characterized by simplicity and less computational intensity, have demonstrated comparable performance. The highest efficiency attained using ChatGPT was a macro-average F1-score of 73%, whereas the method based on BERTimbau-large generated a mean macro-average F1-score of 75%. However, in a cross-dataset scenario, ChatGPT presented better figures (macro-average F1-score of 74% vs. macro-average F1-score of 69%). This comparison underscores the need to consider a balance between computational cost and classification performance in selecting the appropriate model.

2. Methodology

2.1. ChatGPT as hate speech classifier

ChatGPT, developed by OpenAI, is a LLM utilizing the multi-layered Transformers architecture as its information processing units. We believe that the ease of use and accessibility of ChatGPT make its adoption inevitable for natural language processing tasks by non-specialized individuals.

This study employed the OpenAI API with the GPT-3.5-turbo model and the ChatCompletion module to classify social media texts as toxic/hate speech. As the module operates as a chat interface, the inclusion of two types of messages, a system message and a user message, was necessary to create the appropriate prompt. This study explores two

¹<https://chat.openai.com/>

variations of user messages (prompt). One with a more narrow focus, which we refer to as Prompt #1, and another with a broader content message, which we refer to as Prompt #2. Since the queries are in Portuguese, find following the english translation for the reader reference:

System: “Você está realizando análise de texto.” (You are performing text analysis.)

User Prompt #1: “Me diga, com sim ou não, se você considera esse texto tóxico: ” (Tell me, with a simple ‘yes’ or ‘no’, whether you regard this text as toxic.) + text

User Prompt #2: “Responda apenas ‘sim’ ou ‘não’ caso o texto abaixo contenha discurso de ódio, ofensa, agressividade ou outras formas de toxicidade:” (Respond with a simple ‘yes’ or ‘no’ if the following text contains hate speech, offense, aggression, or any other forms of toxicity:) + text

In this case, the “text” refers to each comment (tweet) from the test dataset, which is processed individually. To illustrate, some prompts are structured as follows:

User: “Me diga, com sim ou não, se você considera esse texto tóxico: @user ridículo isso” (Tell me, with a simple ‘yes’ or ‘no’, whether you regard this text as toxic: @user this is ridiculous)

*User: “Me diga, com sim ou não, se você considera esse texto tóxico: @user neguinho tá cansado de apanhar e baixar a cabeça” (Tell me, with a simple ‘yes’ or ‘no’, whether you regard this text as toxic: @user the n****’s tired of taking hits and bowing his head)*

2.2. Baseline methods

For the baseline models, we followed the methodology proposed in [Leite et al. 2020] for BERT-based models since BERT-based models approximate the state-of-the-art for other languages [Zampieri et al. 2019] and Portuguese as well.

The process of classification with BERT-based models encompasses several steps. Initially, the input text undergoes tokenization, breaking it into subwords or WordPieces. Following this, the tokenized input is transformed into high-dimensional continuous representations in a process known as embedding. BERT enhances these token embeddings, integrating token, segment, and positional embeddings to generate a more contextually enriched representation.

The core of the BERT-based model comprises multiple layers of transformer blocks. To facilitate classification, a linear layer is appended atop the BERT model. In this procedure, an activation function is employed - softmax was the chosen function for the work at hand. Lastly, the model undergoes training on labeled data employing an appropriate loss function. In this instance, the binary cross-entropy loss was utilized. In the case of all BERT-based methodologies, the larger variant of the model was employed.

To more faithfully reproduce the work of [Leite et al. 2020], we use the *simpletransformers* library² with arguments set to default values. Three model versions were investigated using different language models: BERTimbau³ [Souza et al. 2020], Distil-

²<https://simpletransformers.ai/>

³<https://huggingface.co/neuralmind/bert-large-portuguese-cased>

Bert⁴ [Sanh et al. 2019], and BertPierreguillou⁵ [Guillou 2021]. In each model, the language base was established on BERT and subjected to training across diverse databases and tasks. The DistilBert is a multilingual case.

A naive approach was also implemented using a non-sequential model for comparison, called Linear Model. Following the preprocessing, the text data was tokenized and transformed into sequences using a straightforward Tokenizer. These sequences were then padded to achieve a consistent length of 280 tokens, corresponding to the maximum length of a tweet, thereby ensuring a uniform input shape for the model. The architecture comprised an Embedding layer with 16-dimensional embeddings, a Flatten layer, a Dense layer containing 32 neurons with a ReLU activation function, and a Dense output layer featuring a single neuron with a sigmoid activation function.

The study focused on binary classification, and the metrics employed included precision, recall, F1-score per class, and macro-F1. For all experiments, data preprocessing consisted of removing links and anonymizing user mentions. All models were trained using binary cross-entropy loss and the Adam optimizer, adhering to standard practices in the field. The source code can be accessed at <https://github.com/ufopcsilab/ToxicSpeech-ChatGPT-STIL>.

2.3. Evaluation Metrics

The models are evaluated with regard to several metrics, including class-specific F1-score, precision, and recall, along with their macro and weighted variants. The “macro” version of these metrics calculates the metric independently for each class and then takes the average, treating all classes equally. At the same time, the “weighted” version calculates metrics for each class independently, but when it averages them, it uses a weight that depends on the number of instances for each class. Confusion matrices are also employed, offering a more visual depiction of model performance across different classes.

3. Results and Discussion

This section presents the datasets used in the experimental design and the results of the experiments aimed at answering the research questions.

3.1. Datasets

The ToLD-Br is the primary dataset employed for evaluating the methodology. In contrast, the HLPHSD is a supplementary dataset aiming for cross-dataset evaluation. Here we only used HLPHSD as a test dataset.

ToLD-Br: The work proposed in [Leite et al. 2020] presents an extensive dataset focused on detecting toxic language in Brazilian Portuguese. Collected from Twitter, the dataset is larger than others found in the literature, covering various demographic groups and considering different types of toxic language: LGBTQ+ phobia, obscenity, insults, racism, misogyny, and xenophobia. Tweets outside of these toxic categories were deemed non-toxic (or non-hateful). A stringent annotation criterion was employed, in which three volunteers independently classified each tweet. A total of 21,000 annotated tweets were

⁴<https://huggingface.co/Davlan/distilbert-base-multilingual-cased-ner-hrl>

⁵<https://huggingface.co/pierreguillou/bert-large-cased-squad-v1.1-portuguese>

selected to compose the dataset, out of which 9,255 were classified as toxic and 11,745 as non-toxic. Data collection took place over two months (July and August 2019). The authors divided the dataset into 80% for training and the remaining for testing using a stratified strategy.

HLPHSD: Proposed in [Fortuna et al. 2019], the HLPHSD consists of 5,668 tweets from 1,156 users collected from January to March 2017. Each tweet was initially labeled in a binary manner (hate vs. no hate) by non-expert volunteers. Subsequently, a second round of labeling was conducted by specialists, with each tweet receiving multiple labels, resulting in a hierarchical taxonomy. In total, 81 hate speech categories were identified. Cohen’s Kappa [Gamer et al. 2012] was used to verify the agreement between annotators. The authors collected data from Brazilian and European profiles, making the dataset diverse regarding this criterion. A total of 31.5% of the tweets are annotated as hate speech.

3.2. Can ChatGPT effectively detect hate/toxic speech in social media texts?

Large Language Models like ChatGP incorporate a certain degree of randomness in their output generation. Thus, the same query to ChatGPT can yield slightly different results. The “temperature” parameter primarily controls this randomness during the inference process. A higher temperature produces more diverse and creative outputs, while a lower temperature leads to more focused and deterministic outputs. To obtain deterministic output from the OpenAI API for each tweet in the ToLD-Br test set, we set the temperature parameter to zero. This implies that every query consistently generates the exact same result. In Table 1, we present the results for the two types of prompts investigated.

Table 1. Classification using ChatGPT-3.5-Turbo on the ToLD-BR test set with GPT Temperature set to zero.

	Prompt #1			Prompt #2		
	Precision	Recall	F1-score	Precision	Recall	F1-score
No-hate	0.84	0.55	0.66	0.80	0.69	0.74
Hate	0.60	0.87	0.71	0.66	0.78	0.72
Macro Avg.	0.72	0.71	0.69	0.73	0.74	0.73
Weight Avg.	0.73	0.69	0.68	0.74	0.73	0.73

3.3. How well does ChatGPT perform compared to models specifically trained for hate/toxicity detection tasks?

The experiment was conducted five times for each baseline approach, altering the seed value to generate variability in the stochastic components. Figure 1 illustrates the comparative results between the baseline method and the classification outcomes achieved by ChatGPT-3.5 Turbo. Each baseline method underwent training for ten epochs, utilizing a learning rate of 3×10^{-5} , batch size of 8, the Adam optimizer, and a binary cross-entropy loss function. Figure 2 displays the confusion matrix for the best model based on BERT and the classification by ChatGPT-3.5 Turbo with Prompt #2.

3.4. How well does the baseline model perform on a cross-dataset evaluation?

In an effort to enhance the complexity of hate speech classification for baseline models, a cross-dataset scenario has been evaluated in this study. The HLPHSD, similar to the ToLD-BR dataset, was collected from Twitter, albeit at a distinct temporal point, and focused

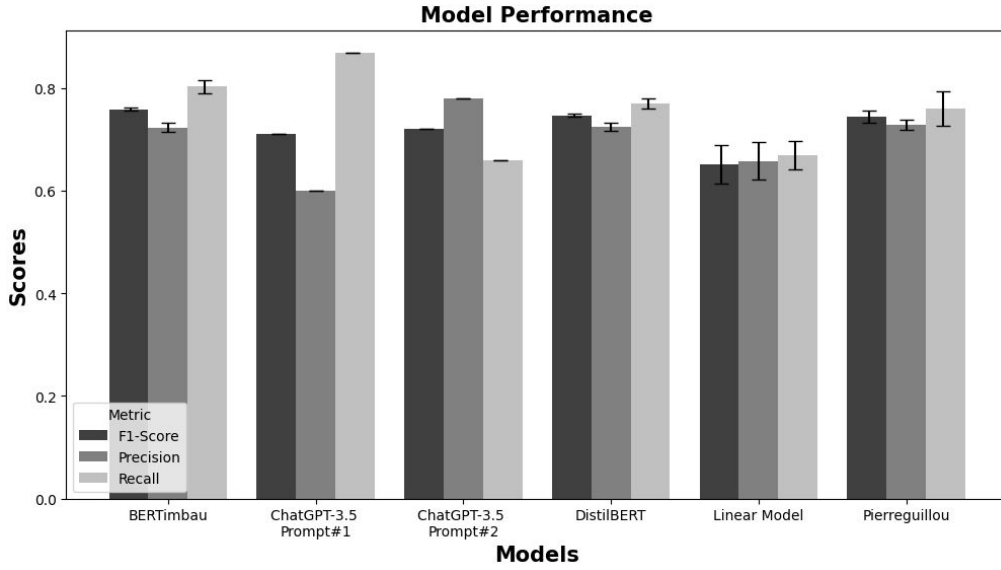
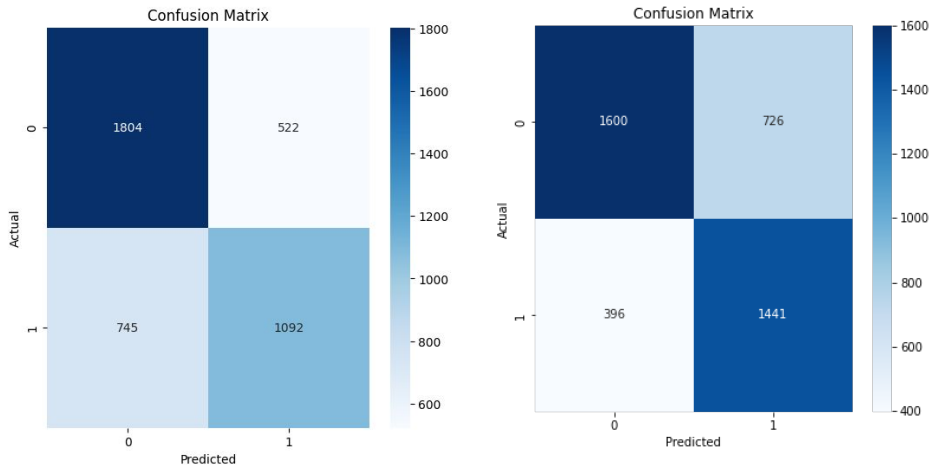


Figure 1. The results for precision, recall, and the F1-score pertaining to the hate speech class (1.0) within the ToLD-BR dataset are presented. Each experiment was conducted five times to ensure reliability.



(a) Confusion matrix for the model based on BERTimbau-large-cased.

(b) Confusion matrix for the ChatGPT-3.5 Turbo classification, with Prompt #2.

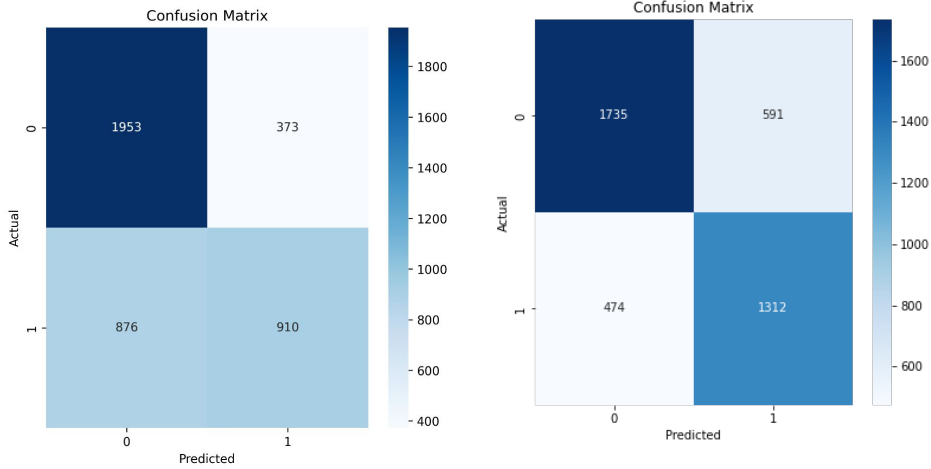
Figure 2. The confusion matrices from an experimental round on the ToLD-Br test dataset.

on Twitter profiles that utilized European Portuguese. The optimal model (employing BERTimbau) trained on the ToLD-BR dataset was selected for cross-evaluation. For this experiment, we balanced the HLPDSD dataset to align the distribution of instances per class with the ToLD-BR test set. The balancing consisted of randomly subsampling the majority class to equal the number of instances in the ToLD-BR test set, while retaining all samples from the minority class (hate speech). The results can be observed in Table 2. Also in Table 2, one can see the performance of ChatGPT-3.5 Turbo on the same dataset, the HLPDSD. Figure 3 displays the confusion matrix for the best model based on BERT

and the classification by ChatGPT-3.5 Turbo with Prompt #2.

Table 2. Cross-dataset results on the HLPHSD-balanced test set in terms of Precision (P), Recall (R) and F1-Score (F1).

	BERTimbau			ChatGPT Prompt#1			ChatGPT Prompt#2		
	P	R	F1	P	R	F1	P	R	F1
No-hate	0.70	0.82	0.76	0.81	0.62	0.70	0.79	0.75	0.77
Hate	0.70	0.55	0.62	0.62	0.81	0.71	0.69	0.73	0.71
Macro Avg.	0.71	0.69	0.69	0.72	0.72	0.71	0.74	0.74	0.74
Weight Avg.	0.71	0.71	0.70	0.73	0.71	0.70	0.74	0.74	0.74



(a) Confusion matrix for the model based on BERTimbau-large-cased.

(b) Confusion matrix for the ChatGPT-3.5 Turbo classification, with Prompt #2.

Figure 3. The confusion matrices from an experimental round on the HLPHSD test dataset.

3.5. Discussion

Figure 1 clearly elucidates that ChatGPT-3.5 Turbo maintains a strong competitive stance, even without the employment of fine-tuning on the training data, hence exemplifying its capability for zero-shot classification. An important observation from the study is that the choice of the prompt dramatically influences the outcome. When the analysis utilizes Prompt #1, narrowly centered around the term 'toxic', it garners an exceptionally high recall albeit with a slight sacrifice in precision. In contrast, Prompt 2, characterized by its wider scope, yields a more harmonious balance between precision and recall metrics, albeit with a slight inclination towards precision. Notably, with the HLPHSD dataset - classified under a hate-speech-oriented taxonomy - Prompt #1 failed to provide any discernible advantages.

Furthermore, ChatGPT's distinct prowess is strikingly accentuated in the context of cross-evaluation, as evidenced by Table 2, as well as in Figure 3. Notably, the BERTimbau-cased-based model, originally trained with the ToLD-BR dataset, exhibited a decrease in performance during cross-evaluation. These findings decisively underline the exceptional abilities of ChatGPT-3.5 Turbo in its role as a Zero-shot Classifier for this task.

An intriguing observation emerged during experiments with ChatGPT, where the model not only classified tweets as “yes” or “no” for toxicity, but also autonomously provided justifications for its decisions, ranging from simple affirmations to sophisticated explanations. This indicates ChatGPT’s understanding of offensive content, which likely contributed to its effective “toxic” categorization. However, its interpretations can skew when offensive language is used colloquially or rhetorically, leading to potential errors. Despite this, ChatGPT’s explanation capability could be invaluable in an industry setting, where model interpretability often holds significant importance.

Regarding computational cost, it is important to note that the ChatGPT-3.5 Turbo model has more than 175 billion parameters. In stark contrast, BERT-based models have an approximate parameter count of 100 million, and the simplest model, referred to here as the linear model, features roughly 1.74 million parameters. The employment of ChatGPT for this task will inevitably necessitate a significant energy outlay, thus entailing greater financial costs for inference.

4. Conclusion

The utilization of LLM-based models, such as ChatGPT, is observing an increasing surge of interest, with expectations pointing towards an expanded deployment in various NLP tasks, inclusive of text toxicity or hatefulness detection by both individuals and corporate entities. The insights gleaned from this study underline the competitiveness of ChatGPT-3.5 Turbo for this task, even when compared to models specifically fine-tuned for the same purpose. The results demonstrate a notable proficiency in toxic speech detection, boasting exceptional performance metrics and showcasing resilience towards shifts in data distribution. However, this prowess comes at a markedly higher computational cost when compared to other models assessed within this investigation. This highlights that, in practical terms, ChatGPT may not be the optimal selection for production use unless the requirement for interpretability is paramount. Smaller language models, such as Falcon [Penedo et al. 2023] and Llama [Touvron et al. 2023], can be an interesting future research path to balance computational cost, performance, and interpretability.

Acknowledgements

The authors would also like to thank *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001*, *Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, grant APQ-01518-21)*, *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant 308400/2022-4)*, and *Universidade Federal de Ouro Preto (UFOP/PROPI)* for supporting the development of the present study.

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.
- Davidson, T., Warmusley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

- de 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.
- Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In Proceedings of the third workshop on abusive language online, pages 94–104.
- Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., and Kendall’s, W. (2012). Package ‘irr’. Various coefficients of interrater reliability and agreement, 22:1–32.
- Guillou, P. (2021). Portuguese bert large cased qa (question answering), finetuned on squad v1.1.
- 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.
- Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th forum for information retrieval evaluation, pages 14–17.
- Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In Proceedings of the first workshop on abusive language online, pages 52–56.
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116.
- Radfar, B., Shivaram, K., and Culotta, A. (2020). Characterizing variation in toxic language by social context. In Proceedings of the international AAAI conference on web and social media, volume 14, pages 959–963.
- Salehabadi, N., Groggel, A., Singhal, M., Roy, S. S., and Nilizadeh, S. (2022). User engagement and the toxicity of tweets. arXiv preprint arXiv:2211.03856.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
- 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.
- 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.
- Vargas, F., Carvalho, I., de Góes, F. R., 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.
- Walther, J. B. (2022). Social media and online hate. Current Opinion in Psychology, 45:101298.
- Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the semeval 2018 shared task on the identification of offensive language. In 14th Conference on Natural Language Processing - KONVENS 2018. Verlag der Österreichischen Akademie der Wissenschaften.
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.

When Tweets Get Viral - A Deep Learning Approach for Stance Analysis of Covid-19 Vaccines Tweets by Brazilian Political Elites

Lorena Guadalupe Barberia¹, Pedro Henrique de Santana Schmalz¹,
Norton Trevisan Roman²

¹Departamento de Ciência Política, Universidade de São Paulo

²Escola de Artes, Ciências e Humanidades, Universidade de São Paulo

{lorenabarberia, pedrosantanaschmalz, norton}@usp.br

Abstract. *Social media platforms are crucial for understanding public opinion about policy issues. In this regard, detecting stance in Twitter posts is a vital tool. In this study, we built a corpus of tweets from 2020 and 2021, annotated with stance towards COVID-19 vaccines and vaccination, and test BERTimbau as a way to automatically detect stance in such tweets. Our model reached 86% accuracy in 2020, 77% in 2021, and 79% in the combined 2020/2021 set. Our results also highlight the time-dependent nature of data distribution and, as a consequence, stance classification. Therefore, this research also contributes to the field by shedding some light on the existing methodological challenges in analyzing complex public policy debates over time.*

1. Introduction

Social media platforms, such as Twitter, are crucial for understanding public opinion on policy issues given their widespread use by the population. Its main drawback, however, lies in this popularity, in the sense that its users create enormous amounts of data every day and analyzing these data can become a cumbersome task. In this regard, automatically detecting stance in such media becomes a vital tool.

Automatic Stance Detection (SD) is a sub-field of Natural Language Processing (NLP) that seeks to automatically identify the stance of the author of some text towards a proposition or pre-chosen target [Mohammad et al. 2017]. Given its increased relevance in public debate and its widespread use by the public, Twitter has received increasing attention as a valuable source for monitoring public opinion (*e.g.* [Walker et al. 2012, Bar-Haim et al. 2017, Dey et al. 2017]).

Understanding the stance of tweet authors on a particular issue provides real-time user-generated information, which is valuable but is also challenging due to the highly nuanced and subjective nature of the language used by those expressing their opinions using this platform. To address these challenges, we introduce a *corpus* of tweets¹ by Brazilian Political Elites (*i.e.* candidates supported by their parties in local elections), collected during 2020 and 2021 and annotated with the position toward COVID-19 vaccines and

¹Available at <https://github.com/PedroSchmalz/covid19-tweets-brazilian-mayoral-candidates> under CC BY-NC-SA 4.0.

vaccination, and propose an approach to automatically classify stance in such tweets².

The remainder of this article is organized as follows. Section 2 presents some related work on stance detection and the use of BERT [Devlin et al. 2018] and BERTimbau [Souza et al. 2020] in NLP tasks. Next, in Section 3 we describe the procedure we followed to collect and classify the *corpus* of tweets, along with the steps taken to train our automatic stance classifier. Section 4, in turn, presents and discusses our results. Finally, our conclusions and directions for future work are presented in Section 5.

2. Related Work

Stance detection has been extensively studied in the field of NLP, and is often used in opinion analysis (e.g. [Bar-Haim et al. 2017]), policy debates (e.g. [Somasundaran and Wiebe 2009, Addawood et al. 2017, Zhang et al. 2017, Augenstein et al. 2016, Walker et al. 2012]), social media monitoring (e.g. [Dey et al. 2017]) and fake news detection (e.g. [Lillie and Middelboe 2019]). We contribute to the stance detection literature by using a deep learning approach to identify political elites’ stances on the subject of COVID-19 vaccines.

In this regard, one of the most successful deep learning models for NLP is the Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al. 2018], which has achieved state-of-the-art results on a wide range of tasks. Minaee *et al.* [Minaee et al. 2021] analyze the increase in the usage of deep learning models for text classification. They review, amongst other things, the rise of Transformer-based PLMs (Pre-trained Language Models). As noted by the authors, when used for sentiment analysis on popular data sets (e.g. IMDB, SST-2, Amazon, etc.), BERT and its variants (BERT-large, RoBERTa, ALBERT, etc.) obtain promising results in classification tasks.

Although BERT has been successfully applied to tweet classification, challenges remain to be addressed. First, it should be noted that there are few *corpora* available in Brazilian Portuguese that could be used to this end. Still, some authors undertake the task of classifying text in Brazilian Portuguese (e.g. [Aguiar et al. 2021, Nascimento et al. 2015, Junqueira and Fernandes 2018, Torres et al. 2020, Brum and Nunes 2017, Silva et al. 2021]) with some of them using BERTimbau [Souza et al. 2020], a pre-trained BERT model for Brazilian Portuguese (e.g. [Martins 2022, Hammes and Freitas 2021, Silva and Freitas 2022]). We seek to contribute to this literature by providing an annotated data set in Portuguese for stance detection, proposing a keyword method for replicating and expanding the *corpora*, and applying BERTimbau to stance classification.

3. Material and Methods

To build our *corpus*, in January 2022 we retrieved tweets from 2020, a year with high uncertainty on the development of COVID-19 vaccines, and 2021, the beginning of the vaccination campaign that changed the debate on COVID-19 immunization from the previous year. Source profiles on Twitter were selected based on candidates registered and certified by the Brazilian Superior Electoral Court (TSE). Of the 300 candidates running

²Replication files available at <https://github.com/PedroSchmalz/when-tweets-get-viral-replication-files>.

for mayor at the 2020 elections in the 26 state capitals, we identified existing Twitter accounts for 243. Among them, 20 profiles were inactive³ and 80 accounts did not publish content related to our research topic during the analyzed period and were therefore excluded from the analysis.

Our final sample consisted then of 143 mayoral candidates. We used Twitter's REST API⁴ to collect all tweets from these candidates and filtered them in using a keyword selection. As noted by [Barbera et al. 2020], this method is preferred over other approaches, such as using subjective categories, because it allows for researcher control and can be replicated and even used in different media.

The set of keywords (Table 1) used in this study was developed in four test trials based on observations of spelling variations, term frequency, and usage. Orthographic and spelling issues were addressed after a preliminary analysis of common variations used by Twitter users. Moreover, we consider both lower and upper cases of terms. Subsequently, we classified tweets that could be excluded after determining that their content, while mentioning keyword terms, did not refer to COVID-19 vaccines or vaccination.

Table 1. Keywords used to retrieve tweets

Topic	Keywords
<i>Vaccines and Vaccination (Gen.)</i>	[Vv]acin; [Vv]assina; [Vv]acinacao [Vv]asina; [Ii]munização; [Ii]munisação; Vaccine Symbol (0)
<i>Covid-19 Vaccines and Laboratories</i>	CoronaVac - [Cc]orona[Vv]ac; [Cc]omuna[Vv]ac; [Ss]inovac AstraZeneca - [Aa]stra[Zz]eneca; [Aa]stra[Zz]enica; [Oo]xford; [Oo]xford; [Oo]xfor; [Vv]axzvria Pfizer - [Pp]fizer; [Pp]eizer; [Pp]pfaizer; [Ff]aizer; [Ff]eizer; [Bb]iontech; [Bb]iontec; [Cc]omyrnaty; [Cc]omimaty; [Bb]iontech [Mm]anufacturing [Gg]mbh; Moderna - [Vv]acina da [Mm]oderna; [Vv]acina [Mm]oderna; [Ss]pikevax; [Mm]oderna [Bb]iontech; mRNA=1273; CX-024414 Sputnik - [Ss]putnik; [Ss]putinik; [Ss]putink; [Ss]putinic; [Ss]putinikV; [Gg]amaleya; [Gg]amaleia Janssen - [Jj]anssen; [Jj]anssen; [Jj]&[Jj]; [Jjohnson & [Jjohnson: [Jj]honson & [Jj]honson: [Jj]onson: [Jj] & [Jj]; [Jjohnson: [Jj]johnsons; [Jj]honson; [Jj]ancen; [Aa]d26.COVS2S Covaxin - [Cc]ovaxin; [Cc]ovachin; [Bb]harat [Bb]iotech Novavax - [Nn]ovavax; [Cc]ovavax; [Nn]uvaxovid; NVX-CoV2373; TAK-019; SARS- CoV-2 rS with Matrix-M1 adjuvant; [Ss]erum [Ii]nstitute of [Ii]ndia; [Nn]ovavax; [Ff]ormulation Sinopharm - [Ss]inopharm; BIBP; [Ss]inofarm Other Laboratories - [Bb]utantan; [Bb]utanta; [Ff]iocruz; @fiocruz; [Ff]iocrus
<i>Localities</i>	[Vv]achina; [Vv]axina; [Vv]acina da [Cc]hina; [Vv]acina [Cc]hinesa; [Vv]acina [Bb]ritânica; [Vv]acina [Cc]ubana; [Vv]acina [Rr]ussia; [Vv]acina da [Rr]ussia
<i>Additional Terms</i>	[Dd]oriavac; [Cc]obaia; [Jj]acaré; [Jj]acare; [Dd]itadória; [Dd]itadoria; [Vv]a-china; [Vv]achinação; [Vv]a-chinacao; [Cc]olateral; [Ee]feito [Cc]olateral; [Oo]brigatoredade; [Oo]brigar

³The activity status of these Twitter accounts was determined manually by coders to ensure these were professional candidate profiles with recent posts.

⁴Documentation available at: <https://developer.twitter.com/en>

As a result, of the 2,335 tweets retrieved in 2020, 1,589 referred to COVID-19 vaccines, which built our 2020 part of the *corpus*. During 2021, we managed to retrieve more than 17,000 tweets, of which we randomly selected a sample of 5,000 and, after the above-mentioned analysis, we kept 4,831 tweets related to Covid-19 vaccines. In total, the *corpus* comprises 6,420 tweets from 2020 and 2021.

Next, the *corpus* was manually annotated by five volunteers, which also reviewed each other’s work. Our unit of analysis was an individual tweet and each post was classified either as *Favorable* to Covid-19 Vaccines, *Neutral*, or *Unfavorable*. Examples of the latter include calling Coronavac “*Vachina*” (vaccine + China), or expressions of doubt about vaccine effectiveness. Favorable tweets relate to those praising vaccines and their arrival. Finally, neutral tweets refer to factual posts, including news posts and those that do not communicate a clear stance about COVID-19 vaccines or vaccination in general.

Table 2 shows the class distribution for 2020, 2021, and the total for both years. During both years, the majority of tweets were classified as favorable. However, there is a higher percentage of unfavorable vaccine tweets in 2020 than in 2021 and a significant increase in neutral tweets in 2021 compared to 2020. In 2021, there was a larger volume of tweets related to Covid-19 vaccines, and most were favorable or neutral.

Table 2. Distribution of Classes

Class	2020	2021	Combined
Favorable	1,319	2,682	4,001
Neutral	82	2,085	2,167
Unfavorable	188	64	252
Total	1,589	4,831	6,420

With this *corpus* at hand, we set out to fine-tune a BERTimbau classifier [Souza et al. 2020] (a pre-trained language model for Brazilian Portuguese, based on BERT [Devlin et al. 2018]) to the task of automatic stance classification. To do so, we first randomly set (with stratified sampling) 10% of the *corpus* apart, as a final test set. We then preprocessed both sets separately, so as to avoid data leakage [Zheng and Casari 2018, Kuhn and Johnson 2019]. Preprocessing consisted of lowercasing, hyperlink removal, and other minor cleanings (retweets, emoticons, etc.).

Next, both *corpora* were tokenized with the BERT Tokenizer (WordPiece algorithm) [Wu et al. 2016, Song et al. 2020]. The models were then trained and validated at the training set (*i.e.* with 90% of the data). Training was done using five epochs, a 32 batch size, and a 2×10^{-5} learning rate for the ADAM optimizer. The model was validated through stratified 5-fold cross-validation [Sobhani 2017], so as to preserve a higher number of observations for the minority classes in all folds.

Three different models were then trained: one with observations only from 2020, another from 2021, and an additional one with the merged sample from both years. The metrics used for model evaluation were training and validation cross-entropy loss (which measures the difference between the output of the model and the expected class) [Sukhbaatar and Fergus 2014], Precision (the fraction of true positives among all positives pointed out by the classifier), Recall (the fraction of all positives the were correctly

identified by the classifier) and F1-score (the harmonic mean between precision and recall) [Olson and Delen 2008].

4. Results and Discussion

Figure 1 shows the average loss in both training and validation sets, along with precision, recall and F1-score for each class, at each epoch and averaged across all folds in 2020. As it turns out, loss values seem to indicate an overfitting process after the second epoch. Figures for precision, recall and F1-score indicate that the model can successfully predict favorable tweets in 2020, peaking at the second or third epochs. At the remaining classes, performance drops, probably as a consequence of the small amount of available data. Overall, two or three epochs are enough for the model to reach stability with this data set.

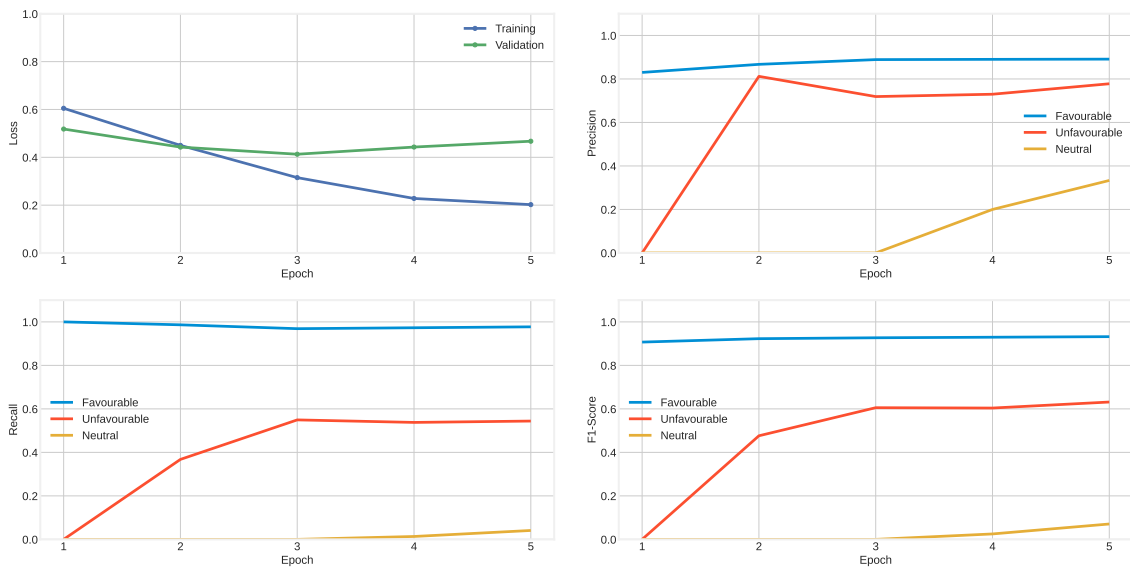


Figure 1. Loss, along with per class precision, recall and F1-score (2020)

During 2021, one sees a raise in the number of tweets related to COVID-19 vaccines and vaccination, possibly due to the authorization and start of vaccination in Brazil. In this year, although Favorable remained as the majority class, minority moved from Neutral to Unfavorable, with a pronounced increase in the Neutral class (see Table 2), thereby changing data distribution. Neutral tweets mostly corresponded to recently elected mayors advertising their municipal vaccination campaign progress.

Figure 2 shows training and validation average loss, along with precision, recall and F1-score for each class, at each epoch and averaged across all folds in 2021. As in 2020, loss values point to an overfitting after the second epoch, with precision, recall and F1-score being higher for the majority class, decreasing as the number of available examples drops, usually stabilizing at the second or third epoch. By comparing results from Figures 1 and 2, one sees how data drift impacts model performance, even for models trained at the specific data set, *i.e.* even after accounting for the new data distribution.

Considering the results obtained from 2020 and 2021, it comes as no surprise that the same behavior can be seen in the union of both sets, as illustrated in Figure 3. In this figure, one sees the model overfitting and all evaluation metrics stabilizing after the

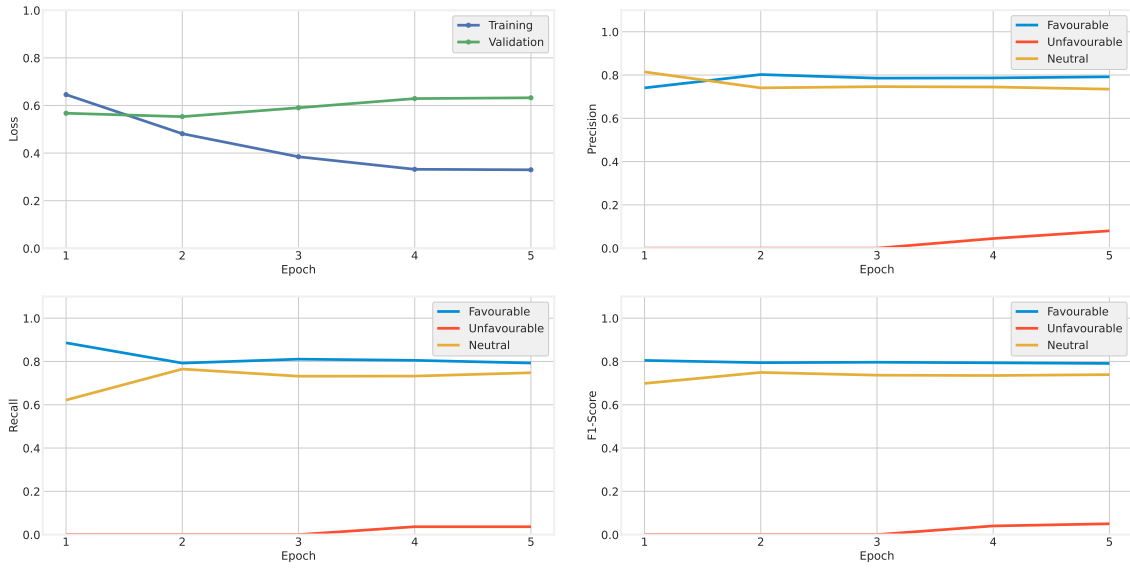


Figure 2. Loss, per class precision, recall and F1-score (2021)

second epoch. Despite this common behavior, overall performance was superior when using data from both years than from each individual year, in the sense that the model was more assertive in its classification of the minority class, even if it still lagged behind the other two classes.

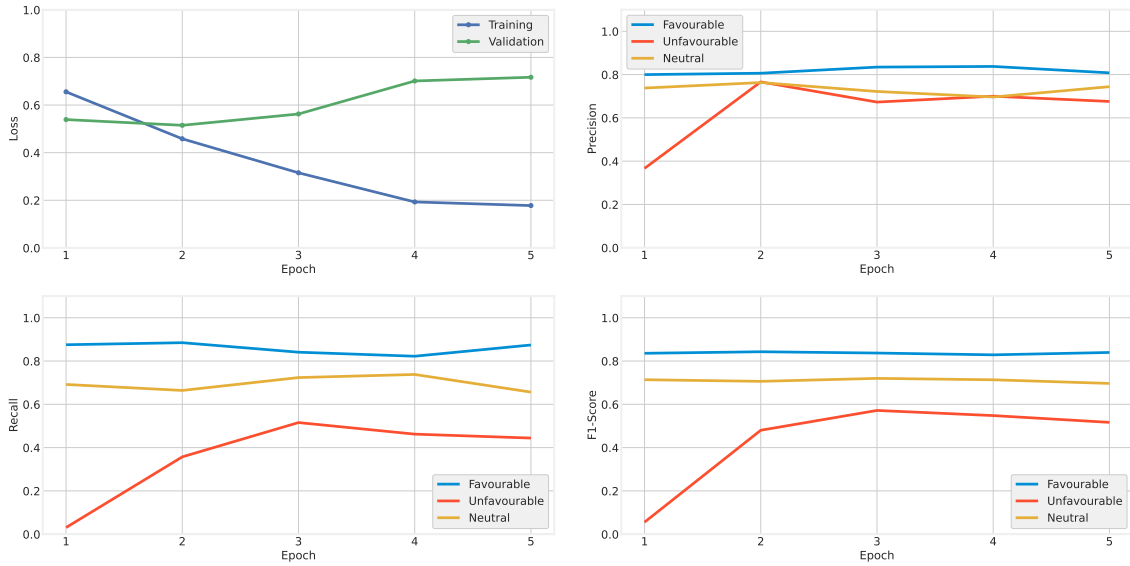


Figure 3. Loss, precision, recall and F1-score (2020 and 2021)

Average validation performance (across the five folds) for all data sets can be seen in Table 3, with a per-class breakdown being shown in Table 4. In these tables, we focus on the results at the second epoch only, so as to avoid overfitting and still reach model stability, and present both average and their associated 95% confidence interval (within parentheses). From these data, it becomes clear the difference in performance between 2020 and 2021 data. Interestingly, when dealing with 2020 and 2021 data together, there was a raise in macro-F1, reflecting the better performance at the minority classes.

Table 3. Average validation results (Epoch 2)

Data Set	Training Loss	Validation Loss	Validation Accuracy	Micro F1	Macro F1
<i>2020</i>	0.45 (0.43:0.46)	0.44 (0.40:0.49)	0.86 (0.84:0.88)	0.86 (0.84:0.88)	0.51 (0.44:0.57)
<i>2021</i>	0.48 (0.48:0.49)	0.55 (0.53:0.58)	0.77 (0.76:0.78)	0.77 (0.76:0.78)	0.67 (0.66:0.69)
<i>2020 and 2021</i>	0.46 (0.45:0.47)	0.52 (0.50:0.53)	0.79 (0.78:0.80)	0.79 (0.78:0.80)	0.67 (0.65:0.69)

Note: 95% Confidence Intervals in parentheses.

Table 4. Per-class average validation results (Epoch 2)

Year	Class	Precision	Recall	F1-Score	Support
<i>2020</i>	Favorable	0.87 (0.856:0.884)	0.986 (0.971:1.00)	0.92 (0.911:0.929)	~237
	Neutral	0 (0.00:0.00)	0 (0.00:0.00)	0 (0.00:0.00)	~15
	Unfavorable	0.81 (0.679:0.941)	0.37 (0.201:0.539)	0.476 (0.303:0.649)	~34
<i>2021</i>	Favorable	0.802 (0.764:0.840)	0.792 (0.735:0.849)	0.794 (0.777:0.811)	~483
	Neutral	0.740 (0.703:0.777)	0.766 (0.702:0.830)	0.748 (0.732:0.764)	~375
	Unfavorable	0 (0.00:0.00)	0 (0.00:0.00)	0 (0.00:0.00)	~12
<i>2020 and 2021</i>	Favorable	0.808 (0.785:0.831)	0.882 (0.85:0.914)	0.840 (0.834:0.846)	720
	Neutral	0.762 (0.716:0.808)	0.664 (0.605:0.723)	0.708 (0.692:0.724)	390
	Unfavorable	0.768 (0.701:0.835)	0.358 (0.297:0.419)	0.480 (0.425:0.535)	~45

Note: 95% Confidence Intervals in brackets.

One possible reason for this difference in class distribution may lie in the observed structural change in the COVID-19 vaccine debate. During 2020, COVID-19 vaccine availability was uncertain, and so candidates tweeted less often and only a few expressed their hope that clinical trials would turn into vaccines. When vaccines were authorized by Anvisa, Brazil's regulatory agency, and began to be offered to some adult groups in January 2021 with a slow roll-out to additional adults, there were significant rise in tweets about COVID-19 vaccines and vaccination. Interestingly, unfavorable tweets were rare in both years, especially in 2021, revealing that these individuals were not likely to voice unfavorable positions on COVID-19 vaccination.

Finally, from the models shown in Tables 3 and 4, the best performance model

Table 5. Results for the Test Set

Year	Class	Precision	Recall	F1-Score	Support
2020	Favorable	0.89	0.98	0.94	132
	Neutral	0.50	0.12	0.20	8
	Unfavorable	0.82	0.47	0.60	19
2021	Favorable	0.79	0.88	0.83	269
	Neutral	0.81	0.71	0.76	209
	Unfavorable	0.0	0.0	0.0	6
2020	Favorable	0.79	0.87	0.83	400
and	Neutral	0.73	0.65	0.69	217
2021	Unfavorable	0.73	0.32	0.44	25

among all folds, in terms of macro F1, was run in the test set, whose results can be seen in Table 5. Results were aligned to those obtained during validation in that the model run on 2020 data performs poorly on the neutral class and the results for unfavorable tweets were worse in 2021. Overall, the results improve over all classes for the combined 2020 and 2021 data set, but there are still problems related to the unfavorable class, with poor recall and F1-scores. In future steps, measures to deal with class imbalance may be explored.

5. Conclusions and Future Work

The present work introduced a *corpus* collected from Twitter with posts from Brazilian political elites about COVID-19 vaccines. The *corpus* was used to detect stance using a deep learning model (BERTimbau). The results indicate that class imbalance increased the difficulty of the task, given the . Poor results were obtained for the minority class in each set. This highlights that even in a short period the model can be affected by data drift, where data distribution changes, influencing the performance of the learner.

This study contributed to the understanding of political discourse during a crisis period (COVID-19 pandemic), which may generalize to other domains or moments. Although BERTimbau is recognized as state of the art for Portuguese, we are still confined to its text representation and the use of a linear layer of neurons as a classifier. In this sense, using its language model with other classifiers could yield better results. This will be a subject of future investigation. Furthermore, the five-fold approach restricts the potential generalization of the model, although it was necessary to maintain class distribution across sets. In future work, we plan to collect additional data to overcome this problem. Finally, we plan to compare our results to those by other commonly used models, following more traditional machine learning techniques, such as support vector machines (SVMs) or random forests (RFs), or other deep learning methods (RNNs, CNNs).

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.

References

- Addawood, A., Schneider, J., and Bashir, M. (2017). Stance classification of twitter debates. In *Proceedings of the 8th International Conference on Social Media & Society - #SMSociety17*. ACM Press.
- Aguiar, A., Silveira, R., Pinheiro, V., Furtado, V., and Neto, J. A. (2021). Text classification in legal documents extracted from lawsuits in brazilian courts. In *Intelligent Systems*, pages 586–600. Springer International Publishing.
- Augenstein, I., Rocktäschel, T., Vlachos, A., and Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.
- Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017). Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.
- Barbera, P., Boydston, A., Linn, S., McMahon, R., and Nagler, J. (2020). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1):19–42.
- Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese (version 1). In *11th International Conference on Language Resources and Evaluation (LREC 2018)*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding (version 2). *arXiv*.
- Dey, K., Shrivastava, R., and Kaushik, S. (2017). Twitter stance detection — a subjectivity and sentiment polarity inspired two-phase approach. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 365–372, New Orleans, LA, USA.
- Hammes, L. O. A. and Freitas, L. A. d. (2021). Utilizando bertimbau para a classificação de emoções em português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, Sociedade Brasileira de Computação.
- Junqueira, K. T. C. and Fernandes, A. M. R. (2018). Análise de sentimento em redes sociais no idioma português com base em mensagens do twitter. In *IX Computer on the Beach*, Brazil.
- Kuhn, M. and Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC.
- Lillie, A. E. and Middelboe, E. R. (2019). Fake news detection using stance classification: A survey (version 1). *arXiv*.
- Martins, G. F. (2022). Um estudo utilizando-se de análise de sentimentos e aprendizado de máquina para a classificação de tweets sobre a vacinação no brasil. Bachelor’s thesis, Universidade Federal do Rio Grande do Sul. In Portuguese.

- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning-based text classification. *ACM Computing Surveys*, 54(3):1–40.
- Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology*, 17(3):1–23.
- Nascimento, P., Osiek, B., and Xexéo, G. (2015). Análise de sentimento de tweets com foco em notícias. *Revista Eletrônica de Sistemas de Informação*, 14(2):2.
- Olson, D. L. and Delen, D. (2008). *Advanced Data Mining Techniques*. Springer, 1st edition.
- Silva, F. and Freitas, L. (2022). Brazilian portuguese hate speech classification using bertimbau. In *International FLAIRS Conference Proceedings*, volume 35. University of Florida George A Smathers Libraries.
- Silva, N. F. F. d., Silva, M. C. R., Pereira, F. S. F., Tarrega, J. P. M., Beinotti, J. V. P., Fonseca, M., Andrade, F. E. d., and de Carvalho, A. C. P. d. L. F. (2021). Evaluating topic models in portuguese political comments about bills from brazil’s chamber of deputies. In *Intelligent Systems*, pages 104–120. Springer International Publishing.
- Sobhani, P. (2017). *Stance Detection and Analysis in Social Media*. PhD thesis, University of Ottawa, Ottawa, Canada. Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment of the requirements for the Ph.D. degree in Computer Science, School of Electrical Engineering and Computer Science, Faculty of Engineering.
- Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore. Association for Computational Linguistics.
- Song, X., Salcianu, A., Song, Y., Dopson, D., and Zhou, D. (2020). Fast wordpiece tokenization. *arXiv*.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417. Springer International Publishing.
- Sukhbaatar, S. and Fergus, R. (2014). Learning from noisy labels with deep neural networks. *arXiv preprint*, arXiv:1406.2080(2(3)):4.
- Torres, B. A., Moraes, L. C., and Pimenta, D. A. C. (2020). Um estudo da aceitação das vacinas contra a covid-19 na rede social twitter utilizando ferramentas de análise sentimental. In *II Workshop de Tecnologia da Fatec Ribeirão Preto*, volume 1-2.
- Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., and King, J. (2012). That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, , Gouws, S., Kato, Y., Kudo, T., Kazawa, H., and ... Dean, J. (2016). Google’s neural

machine translation system: Bridging the gap between human and machine translation. *arXiv*.

Zhang, S., Qiu, L., Chen, F., Zhang, W., Yu, Y., and Elhadad, N. (2017). We make choices we think are going to save us. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press.

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.

The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion

Magali Sanches Duran, Lucelene Lopes, Maria das Graças Volpe Nunes,
Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
{magali.duran, lucelene}@gmail.com, {gracan, taspardo}@icmc.usp.br

***Abstract.** This paper introduces the journalistic portion of the Porttinari treebank, which aims to be a multigenre NLP resource for Brazilian Portuguese. We report the construction of the treebank, in particular, the human-revised portion with 8,418 sentences, whose annotation process lasted almost three years and involved more than a dozen trained annotators. The full treebank offers to the Portuguese-speaking NLP community nearly 4 million sentences annotated according to the Universal Dependencies framework.*

1. Introduction

Treebanks, corpora whose sentences are accompanied with their syntactic trees, play important roles for both linguistic and NLP (Natural Language Processing) research. For linguists, treebanks enable detailed analysis of syntax, leading to the discovery of linguistic patterns and a deeper understanding of language structure. In NLP, treebanks can be used for the development or improvement of computational language processing models, especially those that require labeled data with explicit syntax knowledge.

Among the theoretical options to build treebanks, the Universal Dependencies (UD) project (de Marneffe et al., 2021) stands out as an international grammar framework, with more than 140 languages available in more than 240 corpora. Portuguese has four corpora among the UD datasets: Bosque (Rademaker et al., 2017), PUD (Zeman et al., 2017), CINTIL (Branco et al., 2022), and PetroGold (Sousa et al., 2021). Together, these corpora result in less than a million tokens. In order to contribute on this front, we are committed to the construction of a large multigenre treebank for Portuguese, the Porttinari (which stands for “PORTuguese Treebank”). In this sense, this paper introduces the journalistic portion of Porttinari, which is its first annotated genre, composed by three subcorpora with different characteristics and purposes: **Porttinari-base**, a corpus revised in detail to serve as gold standard; **Porttinari-check**, a small corpus structurally similar to Porttinari-base to serve as testbed and to illustrate the contrast between manual and automatic annotation; and **Porttinari-automatic**, a large corpus that was automatically annotated.

The rest of this paper is organized as follows. In Section 2, we briefly present UD, adopted in the treebank. In Section 3, we present the three subcorpora of Porttinari, and, in Section 4, we draw our conclusions and outline future work.

2. The Universal Dependencies framework

UD (de Marneffe et al., 2021) is a language independent initiative originally designed to annotate morphology, part of speech (PoS) and syntactic dependency relations, in an approach inspired by Tesnière (2015) dependency grammar. UD currently has a fixed set of 17 PoS tags and 37 dependency relations, plus a non-fixed set of morphological features. The UD annotation scheme has enabled the training of automatic classifiers and several comparative studies of language typology. As an example, Figure 1 shows an annotated sentence according to UD. Above the sentence, it is possible to see the relations (labeled arcs), pointing from the heads to their dependents; below them, there are the PoS tags of the words, their lemmas and their morphological features.

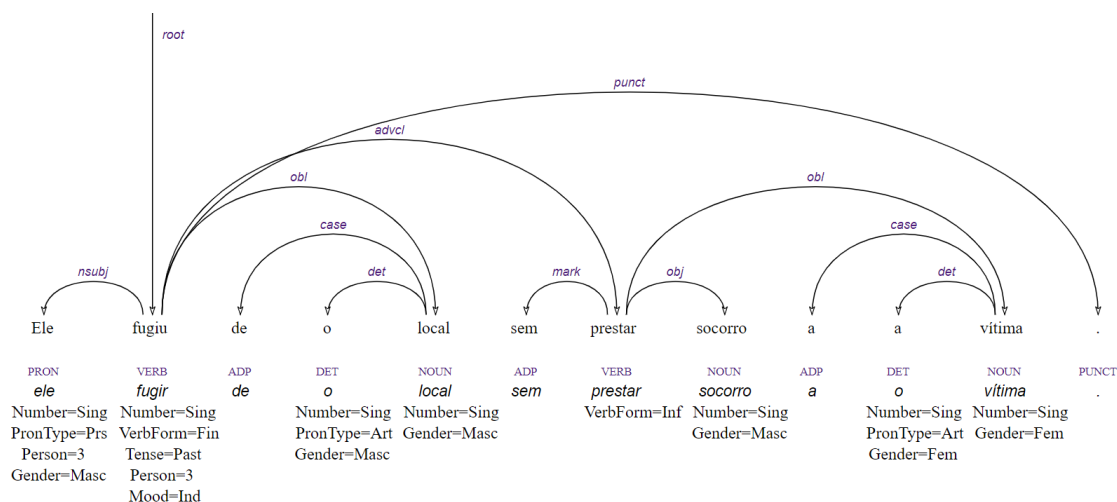


Figure 1. An example of sentence annotated according to UD framework

Much of the annotation in the UD scheme can be accomplished simply by transferring into Portuguese the guidelines described and exemplified in English. However, a number of issues typical of the Portuguese language required linguistic studies and annotation decisions, including phenomena as auxiliary verbs (Duran et al., 2021a), numerals (Duran et al., 2021b) and comparatives (Duran et al., 2023a), among others.

3. The journalistic portion of Porttinari

Porttinari is a large multi-genre treebank (Pardo et al., 2021) and the three subcorpora presented here, containing news texts, are the foundation for the other genres that will follow. Specifically, we used the Folha-Kaggle dataset¹, publicly available, composed of 167,053 news articles extracted from the electronic edition of the Brazilian newspaper Folha de São Paulo published from January 2015 to September 2017.

For preprocessing the corpus, we performed the sentencing and tokenization of each news article. In this phase, Portuguese characteristics had to be taken into account, as the truncated form of infinitive verbs when employed with clitic pronouns (e.g.,

¹ <https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol>

“fazê-lo” was tokenized into the words “fazer” and “lo”) and the contracted words (e.g., “na” was tokenized into the words “em” and “a”). The ambiguities of some contracted words were of particular interest in this part of the process, and we had to use heuristics to solve them. Some examples are the words “consigo” (that can be either an inflection of the verb “conseguir” or a contracted word decomposable into “com” and “si”) and “nos” (that can be a clitic pronoun or a contraction of the words “em” and “os”).

The preprocessing also assigned an ID to each produced sentence with the generic format FOLHA_DOCxxxxxx_SENTxxx, where DOC has a number between 000001 and 167048, referring to the news article the sentence comes from, and SENT has a number referring to the order of the sentence within the news article (no article has more than 999 sentences). For example, the ID FOLHA_DOC006009_SENT013 corresponds to the thirteenth sentence in the 6,009th news article of the original dataset.

Since 5 documents of the original dataset contained no text, the 167,048 news articles resulted in 3,964,292 sentences and 94,799,734 tokens. The three journalistic subcorpora (Porttinari-base, Porttinari-check, and Porttinari-automatic) were produced from this material. Figure 2 depicts the relative size of the three subcorpora in number of sentences. In what follows, we describe each of them.

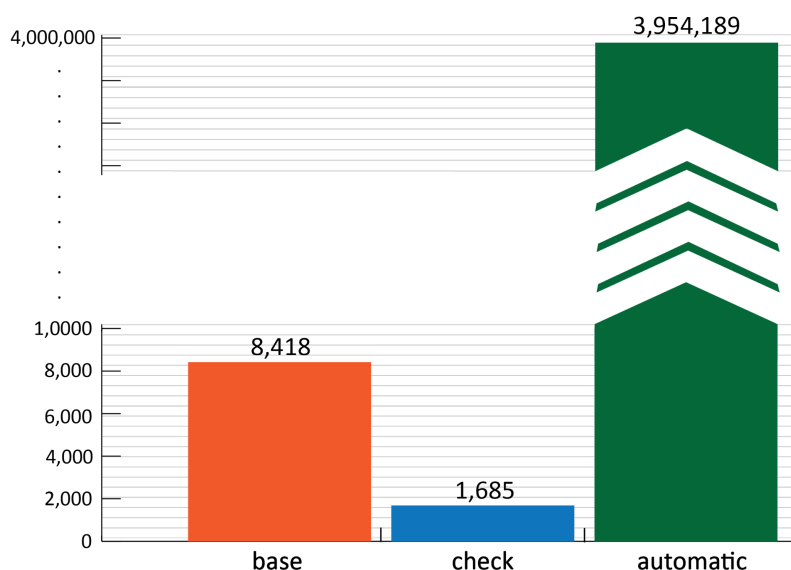


Figure 2. The size of the three subcorpora in terms of number of sentences

3.1. The Porttinari-base subcorpus

The Porttinari-base subcorpus is composed of 8,418 sentences (168,080 tokens - in average, 19.97 tokens per sentence) selected from the initial 5,000 news documents of Folha-Kaggle. The original Folha-Kaggle dataset includes topic classification, as world, economy, education, etc. Therefore, we avoided incorporating full news documents to provide a better diversity of authors and subtopics. The specific choice of sentences was made giving preference for sentences of sizes from 10 to 40 tokens, since small sentences are sometimes too simple and other times are just a juxtaposition of words with no syntax, while too large sentences usually repeat patterns of combined clauses,

either by coordination or subordination. Nevertheless, we also included some sentences out of those bounds, for diversity reasons. Figure 3 shows the distribution of sentences per number of tokens. The resulting subcorpus also has a good lexical diversity of verbal predicates as it has 1,576 different verb lemmas (534 verbs appear only once).

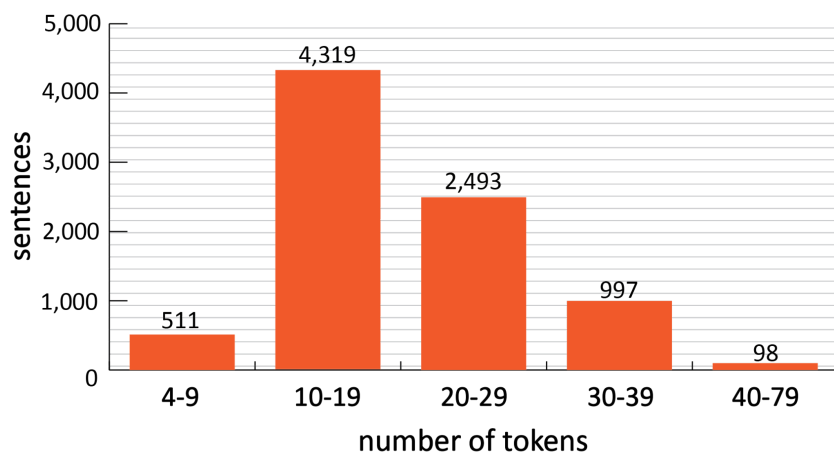


Figure 3. Number of sentences per number of tokens in Porttinari-base

The Porttinari-base annotation process started with an automatic annotation by the parser UDPipe 2 (Straka, 2018) using the Bosque-UD model (Rademaker et al., 2017), which achieves 87% of accuracy (in particular, the Labeled Attachment Score). As the task of revising syntactic trees is complex, we chose to separate the activity into three steps: first we revised the morphosyntactic layer, in which PoS tags are assigned; then we revised the morphological layer, i.e., the lemma and the features of the tokens, in a semi-automated step; and finally we revised the syntactic dependency relations.

The PoS tag manual annotation was performed by trained human annotators, using the Arborator-NILC editor (Miranda and Pardo, 2022) and following strictly the definition made by the PoS tag directives manual (Duran, 2021). The PoS tag distribution achieved is depicted in Figure 4. One of the challenges of PoS tag annotation was the ambiguity of function words, as prepositions and conjunctions, which are sometimes ambiguous with words from other PoS tag classes. Several lexical studies were made to support this task (Lopes et al., 2021; Lopes et al., 2023).

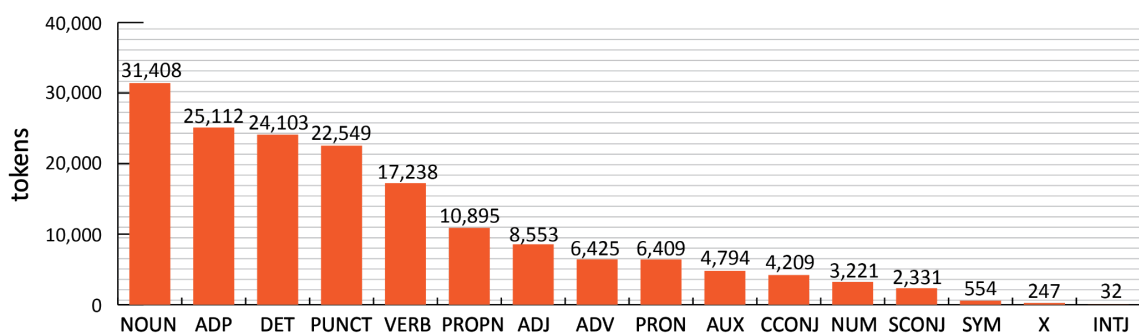


Figure 4. PoS tag distribution in Porttinari-base

The revision of lemmas and morphological features was supported by PortiLexicon-UD (Lopes et al., 2022), which holds the Portuguese forms associated with their PoS tag, lemma, and morphological features using the UD standards (feature names and values). For example, for verbs (either VERB or AUX), the lemma is the infinitive form of the verb and the features include the form of the verb (finite, infinitive, gerund, or participle), and, according to the form, it may include number (singular or plural), gender (feminine or masculine), person (1st, 2nd, or 3rd), mood (indicative, subjunctive, imperative, or conditional), and tense (present, past, future, imperfect, or pluperfect). Human intervention was required to disambiguate cases where more than one combination was possible. For example, the form “*for*” is always a verb, but may have as lemma the verbs “*ir*” or “*ser*”.

The revision of dependencies was a more challenging task. The attribution of dependency relations starts with the definition of the root token, and the other relations are defined from it. Functional words (prepositions, conjunctions, and determiners) usually are only dependents of the relations, while content words (nouns, verbs, adjectives, and adverbs) can be either head or dependent of relations. Some phenomena typical of Portuguese are not predicted in the UD, forcing us to decide how to annotate them consistently throughout the corpus.

The dependency relation annotation becomes an even more complex task when using subrelations, a way to subspecify some of the original 37 UD relations. In our corpora, we adopted 10 subrelations (acl:relcl, aux:pass, ccomp:speech, csubj:outer, csubj:pass, flat:foreign, flat:name, nsubj:outer, nsubj:pass, obl:agent). Given that we do not employ 4 of the original 37 relations (clf, compound, dep, goeswith), we have an overall number of 43 dependency relations distributed as depicted by Figure 5.

The revision of the dependency relations was carefully manually executed over all sentences sequentially, but, due to its natural complexity, we performed an additional vertical revision of several linguistic phenomena. Such a vertical analysis step grants more confidence on the homogeneity of the produced data.

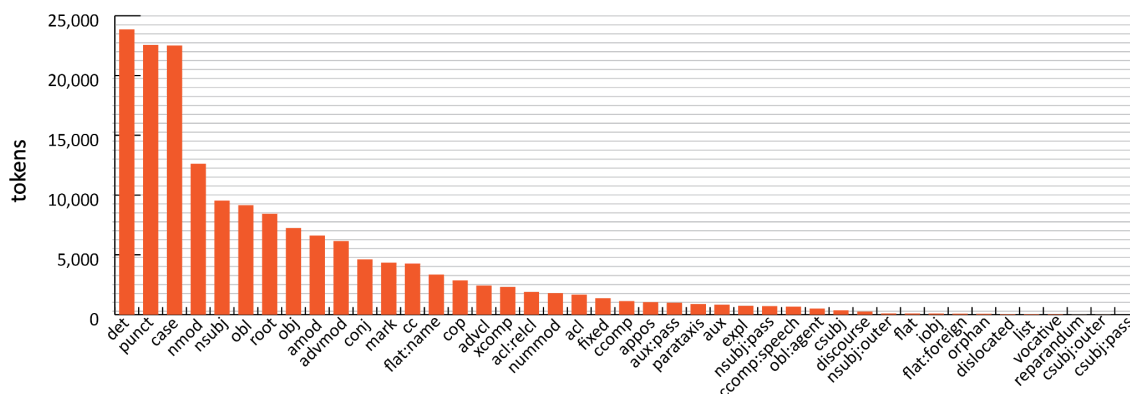


Figure 5. Relation distribution in Porttinari-base

For NLP purposes, Porttinari-base is further subdivided into three subsets: training, validation and test. Such division may be relevant for those interested in training and testing systems using this subcorpus. The subsets have 5,893, 842 and 1,683 sentences in train (70%), dev (10%), and test (20%) files, respectively.

3.2. The Porttinari-check subcorpus

The Porttinari-check subcorpus was developed to provide a controlled dataset to serve as testbed and as a contrast between the careful annotation of the Porttinari-base subcorpus and an automatic annotation. As such, we randomly chose a set of 1,685 sentences (about 20% of the size of Porttinari-base) (consisting of 33,547 tokens, and an average of 19.91 tokens per sentence) with similar characteristics. Specifically, we searched sentences to achieve a proportional distribution of sentence sizes, and similar distributions of PoS tags and dependency relations.

The annotation process of Porttinari-check was fully automatic with UDPipe 2 (Straka, 2018) using Porttinari-base as training set (which, in a preliminary evaluation, showed accuracy results over 98% for PoS tags and 91% for dependency relations).

As a testbed, Porttinari-check was designed to be a complementary evaluation resource. It may be used to check and test (quantitatively or qualitatively) parsing techniques in more varied ways, to complement other evaluation procedures and conclusions (providing varied test sets), to search for and assess specific grammar constructions and to subsidize other studies. As an example, it was already used as the basis for a detailed qualitative evaluation of a parser (Duran et al., 2023b).

3.3. The Porttinari-automatic subcorpus

The annotation of the Porttinari-automatic subcorpus was done entirely automatically, in the same manner as the annotation of Porttinari-check. Although the annotation is completely automatic, careful preprocessing contributes to good results. The quality of the training corpus annotation is critical to ensure that the automatic annotation is consistent.

The purpose of Porttinari-automatic is to produce a very large linguistic resource of NLP for Brazilian Portuguese. It has all the remaining sentences taken from the Folha-Kaggle dataset. As a result, Porttinari-automatic has 3,954,189 sentences, and 94,598,107 tokens, with an average of 23.92 tokens per sentence.

3.4. Overview of distributions in the three subcorpora

We have shown before the distribution of sentence sizes, PoS tags and dependency relations for Porttinari-base only, given it was the seed of the treebank annotation. Now, for comparison purposes, Figures 6, 7 and 8 show such distributions for the three subcorpora.

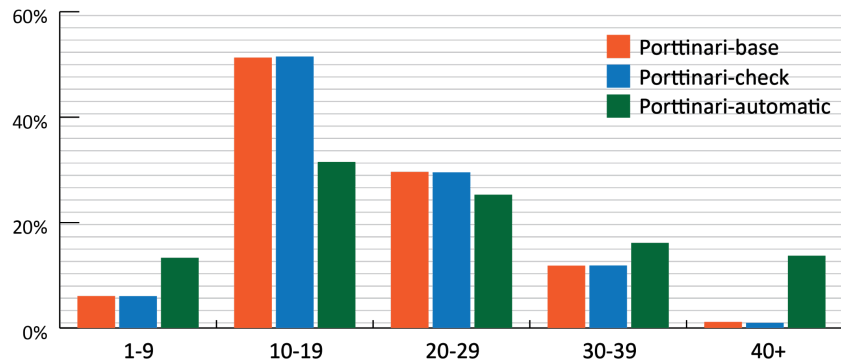


Figure 6. Percentage distribution of sentence sizes in Porttinari-base, Porttinari-check, and Porttinari-automatic

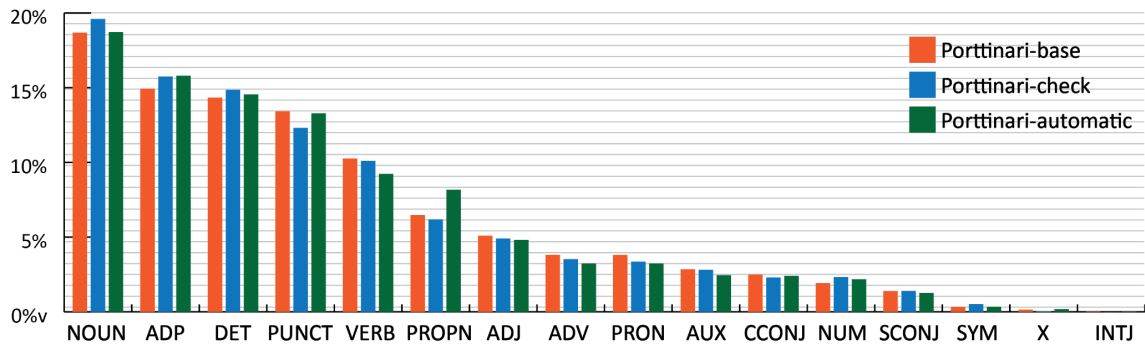


Figure 7. Percentage distribution of PoS tags in Porttinari-base, Porttinari-check, and Porttinari-automatic

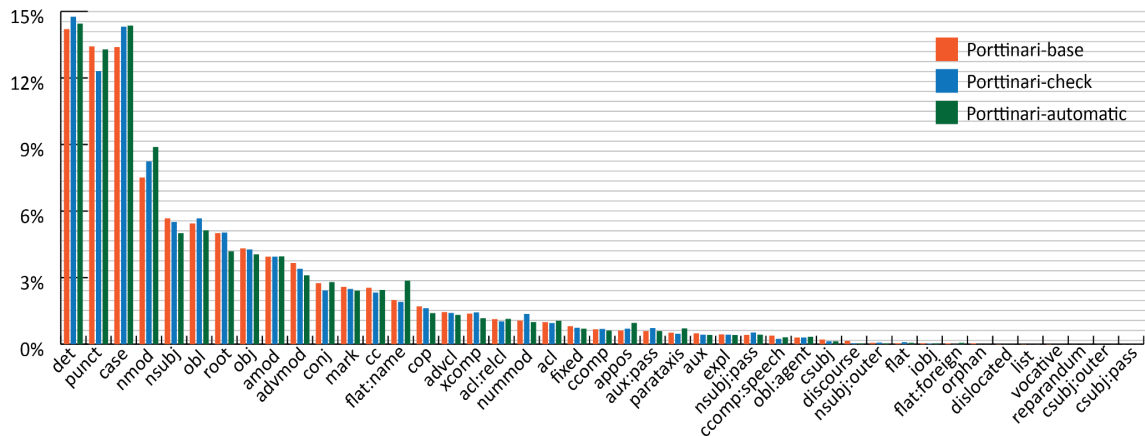


Figure 8. Percentage distribution of dependency relations in Porttinari-base, Porttinari-check, and Porttinari-automatic

It is possible to see that Porttinari-base and Porttinari-check are very similar in their distributions, as Porttinari-check was designed this way. As Porttinari-automatic includes all the remaining sentences of Folha-Kaggle dataset, it diverges from the other subcorpora. Proportionally, it has fewer sentences in the 10-19 size interval and more

sentences in the 40+ interval. Interestingly, it also shows more PROP_N PoS tags, and, consequently, more flat:name relations.

4. Conclusion and Future Work

This paper announces the journalistic genre portion that integrates the large multigenre Porttinari treebank, offering the Portuguese-speaking community a treebank of nearly 4 million sentences annotated with dependency syntax in the UD framework. Together, the three subcorpora that compose the journalistic genre add up to 3,964,292 sentences and 94,799,734 tokens, with an average of 23.91 tokens per sentence. To put this resource in perspective, the Bosque-UD treebank, also for the journalistic genre, has 9,364 sentences and 227,825 tokens, with an average of 24.12 tokens per sentence. The addition of Porttinari data in this scenario places the amount of Portuguese UD annotated resources at the same level of well-resourced languages, which opens several possibilities for NLP applications and linguistic studies.

Future work includes (i) the annotation of enhanced dependencies, which, according to Nivre et al. (2018), have proven to be useful for more advanced applications, (ii) the annotation of semantic roles, following the Propbank model (Palmer et al., 2005), and (iii) the exploration of the treebank for developing NLP research products for Portuguese. In the near future, we also envision to announce a new genre in Porttinari, in particular, tweets, whose annotation process is already advanced.

For the interested reader, Porttinari and related materials are publicly available at the webportal of the POeTiSA project (<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.

References

- Branco, A.; Silva, J.R.; Gomes, L.; Rodrigues, J.R. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In the Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pages 5617-5626.
- de Marneffe, M.-C.; Manning, C.D.; Nivre, J.; Zeman, D. (2021). Universal Dependencies. *Computational Linguistics* 47(2), 255-308.

- Duran, M.S.; Rassi, A.P.; Pagano, A.S.; Pardo, T.A.S. (2021a). On auxiliary verb in Universal Dependencies: untangling the issue and proposing a systematized annotation strategy. In the Proceedings of the Sixth International Conference on Dependency Linguistics (Depling), pages 10-21.
- Duran, M.S.; Lopes, L.; Pardo, T.A.S. (2021b). Descrição de numerais segundo modelo Universal Dependencies e sua anotação no português. In the Proceedings of the VII Workshop on Portuguese Description (JDP), pages 344-352.
- Duran, M.S. (2021). Manual de Anotação de PoS tags: Orientações para anotação de etiquetas morfo sintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do ICMC 434. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Setembro, 55p.
- Duran, M.S.; Lopes, L.; Nunes, M.G.V.; Pardo, T.A.S. (2023a). Construções comparativas em português e sua anotação usando a sintaxe de dependências. Revista da ABRALIN. To appear.
- Duran, M.S.; Nunes, M.G.V.; Pardo, T.A.S. (2023b). Avaliação qualitativa do analisador sintático UDPipe 2 treinado sobre o corpus jornalístico Portinari-base. Relatório Técnico do ICMC 442. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Abril, 58p.
- Nivre, J.; Marongiu, P.; Ginter, F.; Kanerva, J.; Montemagni, S.; Schuster, S.; Simi, M. (2018). Enhancing Universal Dependency Treebanks: A Case Study. In the Proceedings of the Second Workshop on Universal Dependencies, pages 102-107.
- Tesnière, L. (2015). Elements of Structural Syntax. Tradução de OSBORNE, Timothy; KAHANE, Sylvain. Amsterdam: John Benjamins.
- Miranda, L.G.M.; Pardo, T.A.S. (2022). An Improved and Extended Annotation Tool for Universal Dependencies-based Treebank Construction. In the Proceedings of the PROPOR Demonstrations Workshop, pages 1-3.
- Rademaker, A.; Chalub, F.; Real, L.; Freitas, C.; Bick, E.; Paiva, V. (2017). Universal Dependencies for Portuguese. In the Proceedings of the Fourth International Conference on Dependency Linguistics, pages 197-206.
- Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In the Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197-207.
- Lopes, L.; Duran, M.S.; Pardo, T.A.S. (2021). Universal Dependencies-based PoS Tagging Refinement through Linguistic Resources. In the Proceedings of the 10th Brazilian Conference on Intelligent System (BRACIS), pages 601-615.
- Lopes, L.; Duran, M.S.; Fernandes, P.; Pardo, T.A.S. (2022). PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model. In the Proceedings of the 13th Edition of the Language Resources and Evaluation Conference, pages 6635-6643.

- Lopes, L.; Fernandes, P.; Duran, M.S.; Inácio, M.L.; Pardo, T.A.S. (2023). Disambiguation of Universal Dependencies Part-of-Speech Tags of Closed Class Words in Portuguese. In the Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS). To appear.
- Palmer, M.; Gildea, D.; Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics*, 31(1), pages 71-105.
- Pardo, T.A.S.; Duran, M.S.; Lopes, L.; Di Felippo, A.; Roman, N.T.; Nunes, M.G.V. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In the Proceedings of the XIII Symposium in Information and Human Language (STIL), pages 1-10.
- 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.; Kettnerova, V.; Uresova, Z.; Kanerva, J.; Ojala, S.; Missila, A.; Manning, C. D.; Schuster, S.; Reddy, S.; Taji, D.; Habash, N.; Leung, H.; de Marneffe, M.-C.; Sanguinetti, M.; Simi, M.; Kanayama, H.; Paiva, V.; Droганova, K.; Martinez Alonso, H.; Çoltekin, Ç.; 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.; Strnadova, J.; Banerjee, E.; Manurung, R.; Stella, A.; Shimada, A.; Kwak, S.; Mendonca, G.; Lando, T.; Nitisaroj, R.; Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In the Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1-19.
- Souza, E.; Silveira, A.; Cavalcanti, T.; Castro, M.; Freitas, C. (2021). PetroGold – corpus padrão ouro para o domínio do petróleo. In the Proceedings of the XIII Symposium in Information and Human Language (STIL), pages 29-38.

Explorando variações no *tagset* e na anotação *Universal Dependencies* (UD) para Português: Possibilidades e resultados com base no *treebank* PetroGold

Elvis de Souza¹, Cláudia Freitas²

¹Departamento de Letras – PUC-Rio
Lab. Inteligência Computacional Aplicada – PUC-Rio

²Departamento de Letras – PUC-Rio

elvis.desouza99@gmail.com, claudiafreitas@puc-rio.br

Abstract. *The article analyzes variations in PetroGold, a gold standard treebank. The results show that considering the POS tag of multiword expressions in the annotation of all the words that compose them, as well as simplifying the syntactic tagset of the treebank, produces models with better performance on certain metrics, highlighting the importance of linguistic modeling during annotation for adequate natural language processing (NLP) results. The datasets used in the study are available in a dedicated repository and can be further modified to train better language models.*

Resumo. *O artigo analisa variações no PetroGold, um treebank padrão ouro. Os resultados mostram que considerar a classe gramatical das expressões multipalavras na anotação de todas as palavras que as compõem, assim como simplificar o tagset sintático do treebank, produz modelos com melhor desempenho em algumas métricas, destacando a importância da modelagem linguística durante a anotação para resultados adequados no processamento de linguagem natural (PLN). Os datasets utilizados no estudo estão disponíveis em um repositório dedicado, podendo ser ainda mais modificados para treinar melhores modelos de linguagem.*

1. Introdução

Corpora anotados padrão ouro são recursos de extrema relevância no atual cenário do processamento de linguagem natural, em que modelos de aprendizado de máquina podem se beneficiar dos dados para treinar modelos de predição e para avaliar os resultados dos modelos gerados. Por serem “padrão ouro”, há a garantia de que tais recursos passaram por inspeção humana, de tal maneira que as análises linguísticas codificadas na sua anotação são as interpretações humanas dos fenômenos de linguagem. Contudo, o *tagset* e o esquema de anotação de um recurso – quais etiquetas e como serão utilizadas na anotação do *corpus* – pode variar de acordo com os objetivos para os quais o recurso está sendo desenvolvido.

Nesse contexto, partimos de um *corpus* padrão ouro e experimentamos algumas variações no seu *tagset* e na sua anotação com o objetivo de, por um lado, mostrar algumas das muitas possibilidades que um recurso como esse proporciona, fazendo mudanças na anotação que não comprometem a qualidade da informação linguística anotada, e por

outro, mostrar o impacto que essas modificações produzem no aprendizado de máquina, evidenciando o papel da modelagem linguística durante a tarefa de anotação na obtenção de resultados mais adequados para o Processamento de Linguagem Natural (PLN).

O PetroGold é um *treebank* padrão ouro composto por documentos do domínio do petróleo. Foi desenvolvido com o objetivo de gerar bons modelos de anotação morfofssintática, e se insere em um cenário de poucos recursos padrão ouro para português – nenhum especificamente para o domínio do petróleo. Com o amplo uso de grandes modelos de linguagem (LLMs), a relevância de materiais customizados para um domínio e/ou língua fica ainda mais evidente, como mostram [Souza et al. 2020, Lewkowycz et al. 2022, Samuel et al. 2023], o que justifica o desenvolvimento de recursos como o PetroGold.

Embora os resultados apontem para a importância do desenvolvimento do esquema de anotação adequado na produção de bons modelos de linguagem, é importante ressaltar que o foco deste trabalho não é a avaliação dos modelos, mas o papel dos *datasets* no seu treinamento. Por isso, realizamos todos os testes utilizando sempre o mesmo algoritmo e hiperparâmetros, tendo como variável apenas os *datasets* com anotação modificada, que estão sendo disponibilizados em um repositório dedicado¹.

As variações incluem alteração das etiquetas de classes gramaticais para expressões multpalavras, simplificação do *tagset* de anotação sintática e mudança na forma de particionar as frases em conjuntos de treinamento, teste e desenvolvimento para o aprendizado automático. Além disso, tendo como pano de fundo a recém publicação da terceira versão do PetroGold, realizamos também uma breve comparação desta com a versão anterior, colocando em evidência o impacto das revisões linguísticas na geração de modelos de linguagem de melhor qualidade.

2. Versões e variações do PetroGold

O PetroGold é um *treebank* composto por teses e dissertações do domínio do petróleo, com as frases na sequência em que aparecem no texto². O *corpus* contém anotação morfofssintática padrão ouro no formato do projeto *Universal Dependencies* [de Marneffe et al. 2021], uma iniciativa que visa tornar consistente a anotação gramatical em diferentes línguas.

A anotação do PetroGold foi obtida automaticamente, utilizando o anotador Stanza [Qi et al. 2020], e foi revista, na maior parte do tempo, por quatro anotadores familiarizados com a abordagem UD e com a ferramenta de busca, edição e avaliação de *corpora* chamada ET [de Souza and Freitas 2021]. Quando submetidos a um teste de concordância interanotadores [Artstein 2017], os anotadores alcançaram um índice *kappa* de até 95,1%, sugerindo a qualidade dos anotadores e, portanto, da revisão empregada.

O PetroGold é publicado em duas versões: uma versão para o projeto Petrolês, com um *tagset* ligeiramente diferente do *tagset* UD, e uma versão para o projeto *Universal Dependencies*³, que segue as diretivas do projeto e pode ser obtida automaticamente a

¹Disponível em: <https://github.com/alvelvis/petrogold-stil>.

²Do PetroGold foram eliminados apenas elementos como resumo, apêndice e a seção de referências bibliográficas, além de figuras, tabelas e fórmulas matemáticas, uma vez que atrapalham o processamento sintático. Para um detalhamento do material veja-se [de Souza 2023].

³Disponível em: <https://github.com/alvelvis/petrogold-stil>. Acesso em 13 de ago. 2023.

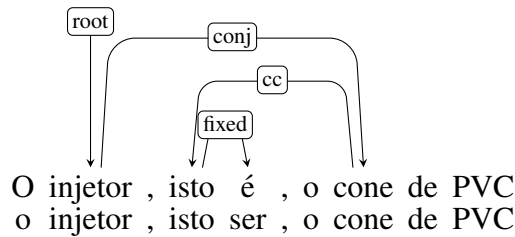


Figura 1. Anotação da MWE “isto é” de acordo com UD

partir da versão Petrolês.

O PetroGold passou por três fases de revisão (versão 1 [de Souza et al. 2021], versão 2 [de Souza and Freitas 2022a] e versão 3, final, apresentada aqui, e cujo processo de construção está detalhadamente descrito em [de Souza 2023]. Esta última versão traz um aumento expressivo no número de revisões da anotação linguística, que foram realizadas para endereçar fenômenos linguísticos anteriormente negligenciados ou para corrigir erros advindos da anotação automática que só puderam ser percebidos com a utilização de novos métodos de revisão semiautomáticos⁴. Entre as modificações realizadas, destacam-se a utilização de um léxico computacional, o PortiLexicon-UD [Lopes et al. 2022], para a revisão de lemas e características morfológicas, além da introdução de três novas etiquetas – *expl:impers*, *expl:pass* e *expl:pv* –, cuja anotação pode ser encontrada em detalhes em [de Souza and Freitas 2023c], e a consolidação de revisões para as expressões multipalavras (MWEs), utilizando três fontes diferentes para obtenção de candidatos a MWE e alinhando os resultados com as diretivas do projeto UD (exemplo na figura 1, e uma descrição sobre a anotação de MWEs no *corpus* pode ser encontrada em [de Souza and Freitas 2023a]).

Em relação ao particionamento das frases dos *datasets* em conjuntos de treinamento, teste e desenvolvimento, as versões Petrolês e UD do PetroGold realizam o procedimento da seguinte forma:

- Petrolês** O particionamento é realizado de forma *aleatória*, sendo o mesmo para as versões 1, 2 e 3 do *corpus*, garantindo que as versões são comparáveis, e seguindo a proporção de 90% de frases para treinamento, 5% para teste e 5% para desenvolvimento⁵.
- UD** O particionamento é realizado *por documento*, de maneira a manter documentos inteiros em cada partição. Assim, as partições de treinamento, teste e desenvolvimento têm, respectivamente, 15 documentos (80% das frases), 2 documentos (12% das frases) e 2 documentos (8% das frases). Embora esta versão não possa ser diretamente comparada com as versões do PetroGold para o projeto Petrolês, o *corpus* pode ser comparado a versões recentes do Bosque-UD, nas quais essa recomendação de particionamento do projeto já é seguida.

⁴Para uma apresentação e avaliação dos métodos de revisão utilizados no desenvolvimento dessa nova versão, ver [Freitas and de Souza 2023, de Souza 2023].

⁵A proporção 90:5:5 para particionamento das frases foi a escolhida para se alinhar ao Bosque-UD v2.8 que, à época, seguia essa proporção e havia sido utilizado como base para comparação da qualidade do *treebank* [de Souza et al. 2021].

A tabela 1 resume as características de cada versão do Petrogold usada neste trabalho, incluindo o número de correções feitas em cada uma delas. O número de correções corresponde ao número de *tokens* que tiveram alguma das anotações linguísticas modificadas desde a anotação automática original. Para enriquecer a comparação, incluímos os dados do Bosque-UD, até agora o único *treebank* revisto em língua portuguesa que integra o acervo UD.

<i>corpus</i>	frases	tokens	correções
PetroGold-v3	8.946	250.605	30.948
PetroGold-ud-2.11	8.946	250.605	N/A
PetroGold-v2	8.949	250.595	21.634
bosque-ud-2.11	9.357	227.827	N/A

Tabela 1. Características dos *corpora*

Cada um dos 4 *datasets* explorados neste trabalho possui também até 3 variações possíveis no *tagset*: variação “base”, variação “mwepos” e variação “simplificado”. A variação “base” é a versão padrão dos *datasets*. No PetroGold v3, corresponde a um *tagset* com 5 etiquetas que nem todos os *corpora* disponíveis no projeto UD possuem. Quatro delas (*obl:arg*, *expl:impers*, *expl:pass* e *expl:pv*) são previstas nas diretivas do projeto UD, embora não sejam obrigatórias, e uma delas (*nmod:appos*) é uma criação nossa, para endereçar alguns fenômenos linguísticos específicos relevantes para o projeto Petrolês⁶, conforme descrito a seguir:

- obl:arg*** Um subtipo da relação *obl* (sintagmas preposicionados dependentes do verbo), exclusivo para quando o sintagma é argumento do verbo. A anotação já foi discutida anteriormente, em [de Souza and Freitas 2022b], e está exemplificada na figura 2.
- subtipos de *expl*** Etiquetas *expl:impers*, *expl:pass* e *expl:pv*, empregadas para especificar o pronome “se” (respectivamente, quando há indeterminação do sujeito, voz passiva sintética e verbo pronominal, frases 1, 2 e 3) [de Souza and Freitas 2023c].
- nmod:appos*** A etiqueta foi criada para anotar os fenômenos da frase 4, em que o termo em negrito não é equivalente ao termo do qual depende sintaticamente, mas tem com ele uma relação não explícita que é facilmente interpretada. A etiqueta é empregada também na anotação da frase 5, em que há uma estrutura de hiperonímia, sendo os termos em negrito hipônimos dos termos do qual dependem, e na anotação da frase 6, na qual há uma referência bibliográfica, sendo que o ano de publicação é anotado como *nmod:appos* dependente do núcleo da referência.
1. ***expl:impers***: A princípio, **trabalhou-se** com a hipótese de que, quanto maior o percentual de esmectita de uma argila, maior seria sua afinidade pelo metal.
 2. ***expl:pass***: Através dos mapas de contorno estrutural do Topo do Embasamento e Topo do Rife **observou-se** a presença de um adensamento das isolinhas na direção NW-se adjacente ao Lineamento Tibagi interpretado na porção continental.
 3. ***expl:pv***: Este estudo **se baseia** nas propriedades magnéticas dos minerais que **se concentram** nas rochas da crosta terrestre.

⁶Quando publicamos o PetroGold no projeto UD, realizamos a simplificação da etiqueta *nmod:appos* para *nmod*, o que está de acordo com as diretivas do projeto.

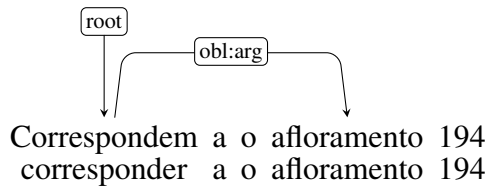


Figura 2. Anotação de argumento verbal introduzido por preposição (*obl:arg*)

4. ***nmod:appos***: Próximo a Presidente Olegário (MG) foram escritos em este estudo pacotes siliciclásticos relativamente espessos (até 60 m) pertencentes a esta formação.
5. ***nmod:appos***: Para fluidos Newtonianos, como a água e o ar, a viscosidade é independente de a taxa de cisalhamento.
6. ***nmod:appos***: A capacidade de absorção de o solvente é proporcional a a pressão parcial em a unidade de absorção (Gupta, 2003).

A variação “mwepos” corresponde ao *dataset* quando damos às expressões multipalavras (MWEs) a anotação de classe gramatical correspondente à classe da expressão como um todo. Assim, na figura que vimos (figura 1), todos os *tokens* da expressão “isto é” recebem a etiqueta de POS “CCONJ” (para conjunção coordenativa), no lugar das etiquetas “PRON” e “AUX”. A solução não é a adotada pelas diretivas do projeto UD, sendo utilizada neste trabalho apenas para evidenciar a possibilidade e comparar os resultados de aprendizado automático com uma opção linguisticamente motivada, apesar de contrária às diretivas do projeto⁷.

A variação “simplificado” corresponde ao *dataset* sem a especificação das etiquetas de relação sintática: *obl:arg* é convertido em *obl*; *expl:{impers,pass,pv}* são convertidos em *expl*, e *nmod:appos* é convertido em *nmod*. Com as versões simplificadas, conseguimos comparar as diferentes versões do PetroGold, uma vez que igualamos os *tagsets*, possibilitando visualizar com maior clareza o impacto das revisões linguísticas nos resultados do aprendizado automático. Essa variação é também a que nos permite comparar o PetroGold ao Bosque-UD, uma vez que este não possui o *tagset* tão especificado quanto o do PetroGold, e a outros corpora em UD que não tenham usado essas etiquetas específicas.

A tabela 2 ilustra as diferenças nos *tagsets* e no particionamento das frases de todos os *datasets*. As células em cinza indicam as características dos *datasets* que os deixam em desacordo com as diretivas do projeto UD.⁸

⁷A informação relativa à classe “geral” das MWEs, codificada como “MWEPOS”, já havia sido adotada – ainda que de forma assistemática e sem revisão – em versões iniciais do corpus Bosque-UD. Anotar MWEs como uma unidade, e não literalmente, como propõe UD, era (e é) a opção de análise do *parser* PALAVRAS [Bick 2014], responsável pela anotação original do Bosque.

⁸Notamos que o *dataset* “bosque-ud-2.11” encontra-se apenas na variação “simplificada” por dois motivos: primeiramente, porque é apenas nesta variação que podemos compará-lo aos outros *datasets*, e em segundo lugar, porque precisamos de fato realizar uma ligeira modificação no *tagset* do *corpus*, uma vez que continha 4 *tokens* anotados com a etiqueta *expl:pass*, sendo simplificados para *expl*.

variação	dataset	<i>obl:arg</i> <i>expl:impers</i> <i>expl:pass</i> <i>expl:pv</i>	<i>nmod:appos</i>	MWEPOS	partições
base	PetroGold-v3	x	x		aleatório
	PetroGold-ud-2.11	x			documento
mwepos	PetroGold-v3	x	x	x	aleatório
	PetroGold-ud-2.11	x		x	documento
simplif.	PetroGold-v3				aleatório
	PetroGold-v2				aleatório
	PetroGold-ud-2.11				documento
	bosque-ud-2.11				documento

Tabela 2. Conteúdo dos *datasets*

3. Metodologia

Para avaliar a qualidade das representações linguísticas codificadas nos *datasets*, usamos tanto uma avaliação intrínseca como uma avaliação extrínseca dos *datasets* [Freitas 2023]. Embora, tradicionalmente, avaliações intrínsecas e extrínsecas sejam usadas para verificar a qualidade de modelos/ferramentas, é possível uma mudança de perspectiva: se na avaliação intrínseca “original” verificamos a capacidade do modelo de generalizar a partir dos dados a que foi exposto, na avaliação intrínseca de *datasets* verificamos (indiretamente) o quanto o *dataset* permitiu esta generalização, levando em conta as características do modelo. A partir dessa mudança de perspectiva, quando olhamos para o desempenho de um modelo, vemos também até onde os dados permitiram ir, considerando os limites do modelo, e pressupondo que (i) o material que serviu de treino está bem anotado e que (ii) o modelo gerado tem um desempenho que não é aleatório. Seguindo com a inversão, a avaliação extrínseca verifica se a informação linguística codificada no *dataset* é adequada para as tarefas mais complexas que o *dataset* pretende auxiliar – o que fazemos quando medimos o impacto das mudanças na codificação de POS das MWEs na anotação de dependências sintáticas. Assim, a avaliação intrínseca de *datasets* anotados verifica a consistência da anotação, e a avaliação extrínseca verifica a adequação de uma anotação para uma determinada tarefas [Freitas 2023].

Os modelos treinados, um para cada *dataset* e variação, são gerados utilizando a ferramenta UDPipe [Straka et al. 2016] na versão 1.2.0, configurada com os parâmetros padrões da ferramenta. As métricas de avaliação intrínseca são as da avaliação conjunta do CoNLL de 2018 [Zeman et al. 2018], com enfoque nos resultados de UPOS (avaliação da anotação de classes gramaticais), LAS (avaliação da anotação da relação e do encaixe das dependências sintáticas) e CLAS (avaliação da anotação da relação e do encaixe das dependências sintáticas considerando apenas palavras de conteúdo lexical).

4. Resultados

A tabela 3 mostra os resultados da avaliação do modelo gerado utilizando cada um dos *datasets* como material de treino. Os números em negrito mostram a variação que treinou o modelo com melhores resultados segundo a métrica daquela coluna (UPOS, LAS ou CLAS) e segundo aquele *dataset* sendo avaliado na linha. Para uma análise

linguística detalhada da qualidade da anotação das relações sintáticas em termos de LAS e CLAS, bem como uma análise de erros, veja-se [de Souza and Freitas 2023b].

<i>dataset</i>	<i>variação</i>	UPOS (%)	LAS (%)	CLAS (%)
PetroGold-v3	base	98,63	89,66	84,66
	mwepos	98,49	89,87	84,79
	simplif.	98,63	90,22	85,61
PetroGold-ud-2.11	base	98,42	88,63	83,30
	mwepos	98,23	89,48	84,33
	simplif.	98,42	89,30	84,38
PetroGold-v2	simplif.	98,40	88,82	83,48
bosque-ud-2.11	simplif.	96,52	81,12	73,51

Tabela 3. Avaliação dos modelos gerados utilizando os diferentes *datasets*

Para os *datasets* “PetroGold-v3” e “PetroGold-ud-2.11”, as variações que produziram melhores resultados de UPOS foram, empatadas, a “base” e a “simplif.”. De LAS, foi a variação “simplif.” para a v3 e “mwepos” para a ud-2.11, e de CLAS, foi a variação “simplif.” para ambos.

O empate entre “base” e “simplif.” na anotação de POS era esperado, uma vez que as variações não contêm nenhuma diferença na anotação de classe gramatical. Em relação à métrica LAS e CLAS, também era esperado que a variação simplificada obtivesse melhores resultados, uma vez que simplificar etiquetas significa necessariamente reduzir o grau de complexidade do que o modelo deve aprender (e para o que deve ser avaliado)⁹. A surpresa, porém, está no fato de que, usando a métrica LAS, os melhores resultados do “PetroGold-ud-2.11” foram da variação “mwepos”, e não “simplif.”, indicando que, embora a simplificação das etiquetas produza números melhores que a versão “base”, nesse caso, modificar o POS das MWEs foi capaz de produzir resultados ainda melhores do que simplificar as etiquetas.

Como esperado, as variações com “mwepos” tiveram desempenho pior que as variações “base” em relação a POS, reforçando que a atribuição de classes de palavras de maneira estática, que não leva em conta o contexto em que as palavras estão inseridas, facilita a generalização das classes. Contudo, para o aprendizado de dependências sintáticas, foco do nosso interesse, vemos uma melhora de até 0,85 p.p. usando a métrica LAS, mostrando que, embora haja perda na anotação de POS, a anotação sintática se beneficia da mudança da classe gramatical das expressões multipalavras, evidenciando o impacto da anotação de um atributo linguístico (classe gramatical) no aprendizado de outro (relação sintática). A mesma tendência ocorre para CLAS, onde a melhora é ainda maior, de até 1,03 ponto percentual.

Considerando que a única diferença entre o “PetroGold-v3” simplificado e o “PetroGold-ud-2.11” simplificado é o modelo de particionamento, já que as etiquetas simplificadas são as mesmas, podemos concluir que, da forma como foram particionados,

⁹Os números devem ser lidos com cautela: embora os melhores resultados sejam os dos *datasets* simplificados, modelos treinados utilizando esses dados não serão capazes, por exemplo, de diferenciar objetos indiretos de adjuntos adverbiais (motivo pelo qual a etiqueta *obl:arg* foi introduzida), de maneira que cabe ao usuário decidir qual *dataset* deseja utilizar no treinamento do seu modelo, conforme seus objetivos.

o *dataset* do projeto Petrolês obteve melhores resultados (85,61% de CLAS) que o do projeto UD (84,38% de CLAS). Embora o particionamento aleatório tenha obtido melhores números, isso não significa necessariamente que seja a melhor forma de particionar um *dataset*, pois é possível que (1) uma outra seleção de frases aleatória obtenha resultados piores, (2) uma outra seleção de documentos por partição obtenha resultados melhores, e (3) a avaliação do modelo considerando frases aleatórias não seja a mais correta, uma vez que somente selecionando documentos inteiros por partição haveria a garantia de que o modelo está sendo confrontado com exemplos de fato inéditos no seu estilo de escrita¹⁰.

Por fim, podemos comparar os *datasets* simplificados em dois grupos: aqueles que têm o particionamento de frases aleatório (modelo do Petrolês) e aqueles que têm o particionamento por documento (modelo do UD). Entre os *datasets* do projeto Petrolês (“PetroGold-v3” e “PetroGold-v2”), vemos que a v3 obtém resultados melhores em todas as métricas em comparação à v2, chegando a até 2,13 p.p. (CLAS) de diferença. Esse é o impacto (positivo) que as revisões linguísticas realizadas nessa nova versão do *treebank* exerceram sobre a geração do modelo de aprendizado automático, considerando que o *tagset* (no caso, simplificado) é o mesmo. Já em relação aos *datasets* que seguem o particionamento do projeto UD (“PetroGold-ud-2.11” e “bosque-ud-2.11”), a diferença é de até 10,87 p.p. (CLAS). Não havendo como confiar em análises especulativas que considerem as diferenças relativas às características dos textos dos *corpora* – o PetroGold é composto por textos do gênero acadêmico, ao passo que o Bosque é composto por textos jornalísticos –, os resultados apenas sugerem que a diferença no desempenho se deve aos vários lotes de revisão pelos quais o PetroGold passou ao longo do tempo, possibilitando uma anotação com maior consistência interna e, portanto, mais facilmente generalizável.

5. Considerações finais

Este artigo explorou diferentes variações na anotação de um *corpus*, visando mostrar as possibilidades que um recurso desse tipo pode oferecer e o impacto dessas modificações no aprendizado de máquina. As variações incluíram alterações na anotação de POS para expressões multpalavras, simplificação do *tagset* de anotação sintática e diferentes estratégias de particionamento para os conjuntos de treinamento, teste e desenvolvimento. Os resultados mostraram que as variações “MWEPOS” e simplificada apresentaram os melhores desempenhos utilizando algumas das métricas, destacando a importância da modelagem linguística durante a anotação para obter resultados mais adequados no processamento de linguagem natural.

Todos os *datasets* testados neste trabalho estão disponíveis em um repositório dedicado¹¹. A ideia é que, com os *datasets* (ou com as ideias que deram origem à produção desses *datasets*), futuramente seja possível expandir este trabalho com a avaliação de fato da qualidade dos modelos associados aos *datasets*, e não apenas da mudança de números produzida por materiais com variações de *tagset* e de anotação.

¹⁰O argumento é um dos apresentados pelo grupo de UD no endereço: https://github.com/UniversalDependencies/UD_Portuguese-PetroGold/issues/3. Acesso em 28 de maio de 2023.

¹¹Disponível em: <https://github.com/alvelvis/ptrogold-stil>. Acesso em 13 de ago. 2023.

Agradecimentos

Os autores agradecem ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, processo #130495/2021-2), à FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, processo #E-26/202.433/2022) e à ANP (Agência Nacional de Petróleo, Gás Natural e Biocombustíveis, Brasil, associada ao investimento de recursos oriundos das Cláusulas de P,D&I, por meio de Termo de Cooperação entre a Petrobras e a PUC-Rio) pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado. Cláudia Freitas atualmente está vinculada ao ICMC/USP e agradece o apoio do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - <http://c4ai.inova.usp.br/>), 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

- Artstein, R. (2017). Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.
- Bick, E. (2014). PALAVRAS, a constraint grammar-based parsing system for Portuguese. *Working with Portuguese corpora*, pages 279–302.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.
- de Souza, E. (2023). *Construção e avaliação de um treebank padrão ouro*. Mestrado, PUC-Rio.
- 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.
- de Souza, E. and Freitas, C. (2022a). Polishing the gold—how much revision do we need in treebanks? In *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–11.
- de Souza, E. and Freitas, C. (2022b). Still on arguments and adjuncts: the status of the indirect object and the adverbial adjunct relations in Universal Dependencies for Portuguese. In *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–10, Fortaleza, Brazil. Association for Computational Linguistics.
- de Souza, E. and Freitas, C. (2023a). Annotation of fixed multiword expressions (mwes) in a portuguese universal dependencies (ud) treebank: Gathering candidates from three different sources. In *Proceedings of the II Universal Dependencies Brazilian Festival (UDFest-BR)*.
- de Souza, E. and Freitas, C. (2023b). Avaliação da anotação automática de dependências sintáticas. *Revista da ABRALIN*.
- de Souza, E. and Freitas, C. (2023c). Um pronome com muitas funções: Descrição e resultados da anotação do pronome -se em um treebank segundo o esquema universal

- dependencies (ud) para português. In *VIII Jornada de Descrição do Português, STIL 2023*.
- de Souza, E., Silveira, A., Cavalcanti, T., Castro, M. C., 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. SBC.
- 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 –. BPLN.
- Freitas, C. and de Souza, E. (2023). A study on methods for revising dependency tree-banks: In search of gold. *Language Resources and Evaluation*, (no prelo).
- Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. (2022). Solving quantitative reasoning problems with language models.
- Lopes, L., Duran, M. S., Fernandes, P., and Pardo, T. (2022). Portilexicon-ud: a portuguese lexical resource according to universal dependencies model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643.
- 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*.
- Samuel, D., Kutuzov, A., Øvrelid, L., and Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.
- 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.
- 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, 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.

Yauti: A Tool for Morphosyntactic Analysis of Nheengatu within the Universal Dependencies Framework

Leonel Figueiredo de Alencar¹

¹Universidade Federal do Ceará (UFC), Brazil

Av. da Universidade 2683 – 60.020-181 – Fortaleza – CE – Brazil

leonel.de.alencar@ufc.br

Abstract. *This paper reports on Yauti, a rule-based morphosyntactic analyzer for the endangered Brazilian indigenous language Nheengatu. Its goal is to generate analyses in the UD framework’s CoNLL-U format. It has been developed on par with the construction of the Nheengatu treebank of the UD collection. In sentences only consisting of known and unambiguous words, the tool generally delivers good results. It obtained a LAS score of 73.2% in a version of the Nheengatu UD treebank with all 1022 sentences automatically provided with XPOS tags and a special annotation to handle non-lexicalized words.*

1. Introduction

Natural language processing has attained near or state-of-the-art results in Brazil in the last two decades. The focus has been on Portuguese, placing it among the languages with the highest Digital Language Support (DLS) Level [Simons et al. 2022, Eberhard et al. 2023]. Despite that, the enormous linguistic diversity in the country, represented by about 150 indigenous languages still alive, has been practically ignored. Only recently has attention been paid to these languages, with the creation of resources such as treebanks and the implementation of computational analysis tools, e.g., [Galves et al. 2017, da Silva Facundes et al. 2021, Gerardi et al. 2021, Martín Rodríguez et al. 2022]. Worldwide, the computational processing of minority languages has enjoyed a growing interest, also on the part of information technology giants, e.g., [Bapna et al. 2022].

This paper describes a contribution to bridging the digital divide that endangers the survival of minority languages in Brazil. We present Yauti, a tool for morphosyntactic analysis of Nheengatu (Modern Tupí) within the Universal Dependencies framework (henceforth UD) [Nivre et al. 2016, de Marneffe et al. 2021].¹ Nheengatu was the most spoken language in the Brazilian Amazon until the middle of the 19th century [Navarro et al. 2017]. Today it is threatened with extinction [Eberhard et al. 2023]. It is the indigenous language that has been most widely learned across Brazil by language enthusiasts or in revitalization initiatives as a means of affirming ethnic identity. It also stands out for its well-documented history spanning four centuries [Moore et al. 1994, Rodrigues 1996, Freire 2011, Rodrigues and Cabral 2011, Moore 2014].

Notwithstanding its cultural, social, and historical significance, there was no publicly available corpus or computational tool for processing Nheengatu until recently. Accordingly, Nheengatu ranks with a score of 0.07 at the bottom of the DLS

¹Yauti is available at <https://github.com/CompLin/nheengatu>.

scale, in a seemingly infinite distance from Portuguese with 0.96 and English with 1.00. Recent efforts to revert this situation include [de Alencar 2021], who reports on a machine translation prototype and a small semantic treebank for stative sentences. [Alexandre et al. 2021a, Alexandre et al. 2021b] deal with a simplistic part-of-speech tagger and a tagged corpus without any syntactic disambiguation.

Yauti represents a much more ambitious initiative. POS taggers, parsers, and annotated corpora are of vital importance not only for developing language technology applications, but also for language documentation, language learning and instruction, and typological investigation. A by-product of the development of Yauti is the UD_Nheengatu-CompLin treebank (henceforth UDTB), the second largest of a Brazilian indigenous language in the UD collection. The tool enabled the treebank to grow relatively fast from 196 sentences in Version 2.11 of 11/15/22 to 860 in Version 2.12 of 5/15/23. As of writing this paper, the development version of UDTB features 1022 sentences.

Section 2 deals with the annotation task according to the UD scheme. Section 3 describes Yauti’s different components and explains how it performs sentence annotations. After reporting on the evaluation results in Section 4, we point out in Section 5 directions for further research.

2. The Annotation Task

A UD treebank’s annotation scheme can be divided into two components. While the first comprises principles considered universal and that every language in the collection must adhere to, the second consists of specific requirements for a particular treebank of a particular language. Both incorporate specifications related to the different levels of morphosyntactic analysis, from orthographic word and syntactic word definitions to parts of speech, feature structures, and syntactic relations. In addition, the treebank file(s) must conform to the CoNLL-U format. A validation program verifies compliance with all of these requirements and grants valid status to a treebank only if it does not violate any of these specifications.²

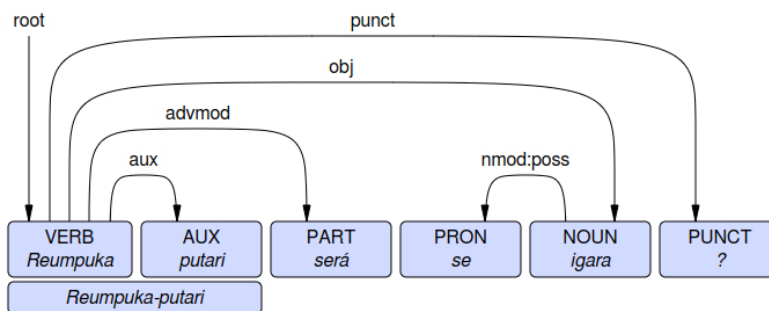


Figure 1. Dependency graph of example (3)

Some of these well-formed conditions are common to different syntactic frameworks, such as the limitation of a single subject per clause. Other requirements seem theory-specific, for example, the prohibition that nouns or pronouns function as adverbial modifiers. UD establishes limited inventories of parts of speech (UPOS) and syntactic relations (DEPREL), which a given language need not exhaust.

²<https://github.com/UniversalDependencies/tools>

Generating an analysis for a Nheengatu sentence conforming to the annotation scheme of UDTB requires filling in nine of the ten fields specified in the CoNLL-U format³ with lexical-grammatical information for each syntactic word; see Figures 1 and 2.⁴ UDTB does not specify yet enhanced dependency relations [Schuster and Manning 2016], housed in column #9.

1	2	3	4	5	6	7	8	10
ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	MISC
1-2	Reumpuka-putari	umpuka	VERB	V	Number=Sing Person=2 VerbForm=Fin	0	root	TokR=0:15
1	putari	putari	AUX	AUXN	Compound=Yes VerbForm=Inf	1	aux	—
2	será	será	PART	PQ	PartType=Int	1	advmod	TokR=17:21
4	se	se	PRON	PRON2	Case=Gen Number=Sing Person=1 Poss=Yes PronType=Prs	5	nmod:poss	TokR=22:24
5	igara	igara	NOUN	N	Number=Sing	1	obj	SpAf=No TokR=25:30
6	?	?	PUNCT	PUNCT	—	1	punct	SpAf=No TokR=30:31

Figure 2. Example of annotation in the CoNLL-U format automatically generated by Yauti for example (3).

Table 1⁵ provides basic statistics of UDTB and MYTH (see Section 4) provided by the `conllu-stats.pl` script from the UD project tools repository. They allow one to dimension the complexity of the task of annotating a sentence in Nheengatu according to the annotation scheme of UDTB.

Table 1. Statistics of UDTB and MYTH

Dataset	Sents	Tokens	Words	Fused	Lemmas	Forms	Fusions	Tags	Feats	Deps
UDTB	1022	10082	10181	99	1075	1450	72	15	66	36
MYTH	43	580	583	3	154	175	3	13	46	24

The first annotation subtask is to determine the syntactic words. Within the context of UD, tokenization requires not just splitting sentences on orthographic words and punctuation. In the cases computed on the Fused column of Table 1, a single token maps onto two different syntactic words. In (1)-(3) from UDTB, the fused words are in bold-face.

- (1) *Amañtu*. ‘I’m just looking.’ (Avila2021:0:0:195)
- (2) *Setá reté pirá paraname*. ‘There are many fish in the river.’ (Avila2021:0:0:341)
- (3) *Reumpuka-putari será se igara?* ‘Do you want to break my canoe?’ (Avila2021:0:0:498)

In (1) and (2), the adverb and postposition cliticize to their heads. In (3), the auxiliary incorporates into the main verb. Yauti handles both cases through a two-step process. First, it splits tokens on white space and separates punctuation. Second, it traverses the

³<https://universaldependencies.org/format.html>

⁴Figure 1 was generated from the CoNLL-U annotation in UDTB by the viewer at <https://urd2.let.rug.nl/~kleiweg/conllu/>. For presentation purposes, we manually edited the program output displayed in Figure 2. *TokR* and *SpAf* are abbreviations for *TokenRange* and *SpaceAfter*, respectively.

⁵*Sents* and *Deps* refer to the number of sentences and syntactic relations. The values from the Lemmas column onwards exclude repetitions.

resulting list of strings and splits off each of a predefined list of suffixes, signaling it with a specific mark to indicate whether it is a compound member or a clitic.

Yauti, using the CoNLL-U Parser⁶ Python library, generates an object of the `Token` class for each orthographic word and each syntactic word. The identity of each `Token` object is an integer or a range of integers, as in Figure 2. Each syntactic word must receive a label from the UPOS and XPOS inventories. The feature structures (FEATS) of each syntactic word must be assigned. The head of the word and the type of syntactic relation (DEPREL) it maintains with that head must be determined.

This last subtask gets especially tricky in complex sentences with coordinated and/or subordinated clauses, clausal complements, etc. Commas, dashes, semicolons, colons, and quotation marks attach to the highest node of these constituents. Failure to identify this highest node triggers the wrong annotation of its dependent punctuation marks. Currently, UDTB has 1913 punctuation marks, totaling 13 different types.

The last column of the CoNLL-U format specifies, via the *TokenRange* attribute, the span of the syntactic word and, through the *SpaceAfter* attribute, whether it is followed by white space or not. As the first situation is the most general, this attribute is only used when the value is negative. Its absence implies a positive value. Other attributes housed in this last column are *Orig* and *OrigLang*, which specify, respectively, the original form and the language of origin of loanwords not yet lexicalized in Nheengatu.

3. Program Construction and Operation

It is evident from the examples presented above that annotating sentences according to the UDTB scheme is a complex task. Therefore, we decompose it into a series of subtasks, performed by different functions implemented in a series of Python modules.

Yauti's starting point is a glossary in JSON format with 1552 entries, generated from a raw text file that a linguist without programming skills can easily edit; see example entry in (4). The first version of this glossary only contained about 850 entries. We created it by extracting the raw text from [de Almeida Navarro 2016]'s glossary, after some post-editing to remove page numbers, correct inconsistent formatting, etc. The circa 700 new entries are a manual sample from the over 8000 entries in [Avila 2021].

- (4) *maã* 5 (*verb*) - to see, to look

Navarro's word classification underwent profound changes to handle annotation in the UD scheme. Instead of his original system of 12 classes and 6 pronoun subclasses, Yauti, based, e.g., on [da Cruz 2011] and [Avila 2021], adopts a much more granular inventory with 83 XPOS tags. For example, instead of a single tag for all particles, Yauti has 20 tags for different particle types. Particles play a fundamental role in Nheengatu, whose verb inflection is relatively poor. Pronouns, determiners, and adverbs also have very high granular labels. Pronouns and determiners are divided into 15 subcategories, e.g., INDQ for indefinite quantitative pronouns. Adverbs are classified into 22 subcategories, such as ADVT, ADVCL, and ADVS, for temporal, locative, and intensity adverbs.

The glossary in JSON format feeds a rule-based morphological generator, con-

⁶<https://pypi.org/project/conllu/>

sisting of functions that take as parameters lemma, part of speech, and inflectional class information, deriving conjugated verbs, pluralized nouns and pronouns, and words with relational prefixes.⁷ This generator populates a full-form lexicon of Nheengatu in the form of a Python dictionary mapping word forms to a list of [lemma, tags] pairs, where tags is a sequence of tags separated by the plus sign. For example, the word forms *amaã* ‘I see’ and *reumpuka* ‘you break’ from (1) and (3) map to [['maã', 'V+1+SG']] and [['umpuka', 'V+2+SG']], respectively. The ambiguous word form *setá*, which functions as a third class verb (V3) in (2), maps to [['setá', 'ADVS'], ['setá', 'INDQ'], ['setá', 'V3+NCONT']].

Using the information encoded in the lexicon, Yauti performs an initial filling of different annotation fields of the `Token` objects, notably UPOS and FEATS. It assigns the feature `Clitic=Yes` or `Compound=Yes` to suffixes that constitute syntactic words, as in (1)-(3). For some word classes, such as ADP, INTJ, CCONJ, and SCONJ, the syntactic relation is already specified at this stage. Additionally, Yauti parses productive morphological derivations. One noteworthy aspect of Nheengatu is the application of aspectual suffixes not only to auxiliary and main verbs, but also to nouns, adverbs, etc.

Yauti then inserts each `Token` object in a `TokenList`. Then it traverses this list token by token, executing functions that fill in the head and syntactic relations, based on available information from the current token and preceding and succeeding tokens.

To this end, Yauti makes use of general syntactic patterns of Nheengatu. Let’s look at some examples. A noun that does not govern an adposition and precedes a verb generally functions as its subject, while an analogous noun following a verb functions as its object. The closest noun to the left of a postposition constitutes its head. This noun, in turn, is linked to the nearest verb through the oblique syntactic relation, noun-dependent adpositional phrases being rare. In noun sequences, the one farthest to the left is a possessive modifier of its next neighbor to the right, which, in turn, is a possessive modifier of the next noun in the sequence, and so on.

Subordinating conjunctions (SCONJ) behave analogously to postpositions: the closest verb to the left is its head, and this verb is linked as an adverbial clause modifier to another verb. Yauti adopts the following heuristics to identify the head of a verb governing a SCONJ: try to find a verb in a preceding matrix clause, if there is none, try to find one in a subsequent clause. Under the scope of the negation particle, however, subordinating conjunctions and the relative pronoun precede their governing verb.

Very typical of Nheengatu are clauses in parataxis, where the first verb governs the second, which governs the next one, and so on. Yauti often manages to handle this construction. However, it fails to correctly analyze clausal complements (*ccomp* and *xcomp* in UD) since Nheengatu has no complementizers or infinitives, and information on verb valence is presently unavailable. Another still unsolved challenge is correctly identifying the head of a noun between two verbs linked via parataxis. It can either be the *obj* of the first verb or the *nsubj* of the second. Yauti always chooses the latter option, which is sometimes incorrect.

As there are no comprehensive formalized descriptions of Nheengatu, the con-

⁷Relational prefixes encode, e.g., syntactic contiguity (CONT) or non-contiguity (NCONT) between a head and its dependent [Rodrigues and Cabral 2011].

struction of the syntactic annotation algorithm proceeded incrementally. We initially made a very simple version of the algorithm to handle basic cases like the SVO pattern in simple sentences. We then applied it to increasingly complex and diverse examples, observing the mistakes made and correcting the code to avoid them. The current algorithm is the result of this process after the annotation of more than 1000 sentences (Table 1). The vocative is an example of a syntactic relation whose annotation has only recently become automated. Before that, Yauti incorrectly assigned the *nsubj* or *obj* relation to verb dependents in the vocative function, following the most general SVO pattern. Now a rule corrects this initial annotation, provided some conditions are met.

```
>>> s='Setá/v3 reté/adv pirá paraname.'
>>> tk=AnnotateConllu.parseSentence(s)
>>> print(tk.serialize())
1      Setá      setá      VERB      V3      Number=Sing|Person=3|Rel=NCont|VerbForm=Fin      0      root      _
enRange=0:4
2      reté      reté      ADV      ADVS      AdvType=Deg      1      advmod      _      TokenRange=5:9
3      pirá      pirá      NOUN      N      Number=Sing      1      obj      _      TokenRange=10:14
4-5     paraname
4      paranã     paranã     NOUN      N      Number=Sing      1      obl      _      SpaceAfter=No|TokenRange=15:24
5      me      upé      ADP      ADP      Clitic=Yes      4      case      _
6      .      .      PUNCT      PUNCT      _      1      punct      _      SpaceAfter=No|TokenRange=25:26
```

Figure 3. Parsing example (2) with Yauti in the Python IDLE shell

Yauti’s central function is `parseSentence`, which takes as an argument a string with or without part-of-speech tags and other abbreviations, as explained below. This function returns an object of the `TokenList` class. Thus, the program can be used in batch processing as a component of a pipeline, generating analyses in the CoNLL-U format, following the UDTB annotation scheme for the sentences given as input.

However, because of ambiguity, the results generated by the fully automatic mode of the program are in general precarious. For each syntactic word with n XPOS tags, the program generates n `Token` objects. Thereafter, `TokenList` creation rules operate on these objects, generating chain errors. To address this problem, an automatic part-of-speech tag disambiguator is under development. For the time being, it is in interactive mode that the program is most effective. In a typical interaction, the user types in an example in the Python shell, executing the `parseSentence` function. Yauti generates the respective analysis. The user checks which words are ambiguous and manually eliminates the ambiguities by specifying the corresponding tag, as in Figure 3. Next, the user reruns `parseSentence` on the tagged sentence, manually correcting detected errors, e.g., after inspecting the visualization of the sentence or executing the validation program. In Figure 3, Yauti incorrectly assigned node 3 the *obj* instead of the *nsubj* relation.

In addition to disambiguating tags, `parseSentence` accepts two other types of abbreviations. First, @ assigns a word the *root* syntactic relation. This generally improves the annotation of verbless sentences, which are common in Nheengatu. They can be tricky for Yauti to deal with, as any content word can be the main predicate of the sentence and act as *root*. Second, we have functions for parsing words the morphological analyzer does not know about and are inappropriate for Yauti’s lexical database. These unknown words fall into two groups: (i) Portuguese words in code-switching, proper nouns, and non-lexicalized interjections, e.g., onomatopoeias; (ii) productive morphological derivations, e.g., collective nouns, words with degree, privative or aspectual suffixes, etc.

4. Evaluation of Yauti

UDTB is the only Nheengatu syntactic treebank available. To evaluate Yauti, therefore, we could not count on a previous gold standard. Instead, to assess how useful the tool can be for annotating texts in Nheengatu, we resorted first to UDTB (Table 1). Given the grammatical and lexical diversity of this treebank, compiled from excerpts of different genres from different regions and diachronic stages, we expect that the average performance in this dataset should reflect in the annotation of other sentences from the same sources or other texts of more or less similar characteristics.

Parsing Nheengatu is especially challenging due to the lack of spelling standardization and limited computational resources. On the other hand, the current lexicon of the morphological analyzer only covers a fraction of [Avila 2021]. Another still unresolved issue is the ambiguity resolution. Thus, we have restricted the evaluation task to three main dimensions by isolating the effects of the unknown word analysis and ambiguity resolution subtasks. The first dimension is the performance in the standard LAS and UAS metrics [Straka et al. 2016, Nivre and Fang 2017, Straka 2018]. The second is the splitting of tokens into syntactic words. The third is morphological analysis, encompassing both lemmatization and UPOS and XPOS assignment. Additionally, we computed accuracy for the *SpaceAfter* attribute.

To perform this evaluation, we implemented a Python code that traverses each `TokenList` of UDTB and extracts, from each `Token` object, a triple in the format (`form`, `tag`, `spaceafter`). The second member of this triple can either be the lower-cased tag from the XPOS field or an abbreviation of the name of a `Token` object construction function, prefixed with = and possibly followed by a series of colon-separated strings in the form of `k|v`, where `k` is a keyword argument and `v` its value. From these triples, another function constructs a valid input for the `parseSentence` function. (5)-(7) are test sentences automatically constructed from the information encoded in UDTB. In (5), `=v` and `=n` trigger the creation of verb and noun `Token` objects for Portuguese loanwords not registered in [Avila 2021]. The function `=hab` in (6) and (7) handles the frequentative aspect suffix *-wara*, which attaches to the verb *sú* ‘to go’ in (11) and to the adverb *iké* ‘here’ in (7). The named parameters `x` and `a` handle the XPOS and the accentuation of the final vowel of the base form.

- (5) *Presizu/nec aintá/pron uistudari/=v português/=n upé/adp*. ‘It’s necessary that they study in Portuguese.’ (MooreFP1994:0:0:8)
- (6) *Asuwara/=hab:x|v:a|t mikití/advc*. ‘I always go there.’ (Avila2021:0:0:432)
- (7) *Ikewara/=hab:x|advdx:a|t aikú/cop, reyuri/v ramé/sconj, resika/v se/pron2 piri/adp*. ‘I’m always here, when you come, come visit me.’ (Avila2021:0:0:429)

Applied to all 1022 thusly annotated sentences from UDTB, Yauti obtained 80.0 and 73.2 in the UAS and LAS metrics, respectively (Table 2). These values are similar to those achieved in some treebanks by the deep neural parser of the UDPipe 2.0 prototype in the CoNLL 2018 UD Shared Task competition [Straka 2018]. For example, this parser, representative of the state-of-the-art in dependency parsing, achieved a UAS of 78.66 and LAS of 74.25 in the Galician-TreeGal treebank. In the other two principal dimensions,

Yauti delivered a much higher performance. Specifically, it achieved an accuracy score of 99.7% and 98.4% in the assignment of lemmas and features, respectively, attaining 100% accuracy in tokenization. It is worth noting that several lemma and feature errors have turned out to be incorrect annotations of UDTB. Yauti also exhibited high accuracy in the *SpaceAfter* attribute, with a score of 99.5%.

Table 2. Performance of Yauti in two data sets

Dataset	LAS	UAS	Lemmas	Feats
UDTB	73.2	80.0	99.7	98.5
MYTH	71.0	76.3	96.6	96.8

This experiment also revealed 96 XPOS and 204 UPOS discrepancies compared to UDTB, many of which were actually incorrect annotations in the treebank. Most errors Yauti commits in this domain involve distinguishing pronouns from determiners, on the one hand, and auxiliary from main verbs, on the other. In the lexicon, these distinctions are underspecified. Yauti tries to guess the correct tags during the construction of the `TokenList` object. However, the accuracy levels show that the algorithm needs improvement in this aspect.

As a first step to test performance on unseen data, we applied the tool to the myth “How the Night Appeared” [de Magalhães 1876, pp. 163-171]. We modernized the spelling but kept the original punctuation intact.⁸ This story contains relatively many obsolete forms. We manually updated the JSON glossary with 34 lemmas hitherto unknown to Yauti. We also manually annotated the sentences, as in (5)-(7). We corrected Yauti’s output and compared this gold standard to the test version. Table 2 displays the main results.

5. Final Remarks

We have encountered several annotation errors we could not yet prevent due to time constraints or a lack of fuller understanding of the respective phenomena. One of the primary issues we face is with verb-subject clauses (Figure 3). Cross-linguistically, this construction relates to unaccusativity, which is difficult to detect automatically. Another challenge comes from verbless sentences, where any content word can act as *root*. A further difficulty is the indirect object (*iobj*), defined in UD theory as a verb’s core argument in addition to an *obj* or a *ccomp*. While identifying it may be challenging in the general case, it seems clear-cut in languages such as English that feature double objects or in languages like German that use a dative case. In UDTB, a noun governing a dative postposition is an *iobj* whenever the same verb governs an *obj* or a *ccomp*. However, as we saw, the latter relation is hard to identify in Nheengatu. We intend to address the issues identified so far soon in the hope of improving Yauti’s performance.

The evaluation experiment showed that Yauti is useful not only to annotate new sentences but also to check existing annotations for consistency. As UDTB constantly increases, we hope to train a neural parser on the data with UDPipe 2.0 and compare its performance with the improved rule-based approach.

⁸We owe the orthography adaptation to the philological expertise of Marcel Twardowsky Avila (p.c.).

References

- Alexandre, D. M., Gurgel, J. L., and de A. Araripe, L. F. (2021a). Compilação de um corpus etiquetado da Língua Geral Amazônica. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 427–431, Porto Alegre, RS, Brasil. SBC.
- Alexandre, D. M., Gurgel, J. L., and de Alencar Araripe, L. F. (2021b). Nheentiquetador: Um etiquetador morfossintático para o sintagma nominal do nheengatu. *Revista Encontros Universitários da UFC*, 6:1–13.
- Avila, M. T. (2021). *Proposta de dicionário nheengatu-português*. PhD thesis, Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.
- Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., Breiner, T., Axelrod, V., Riesa, J., Cao, Y., Chen, M. X., Macherey, K., Krikun, M., Wang, P., Gutkin, A., Shah, A., Huang, Y., Chen, Z., Wu, Y., and Hughes, M. (2022). Building machine translation systems for the next thousand languages. Technical report, Google Research.
- da Cruz, A. (2011). *Fonologia e gramática do nheengatú: A língua falada pelos povos Baré, Warekena e Baniwa*. LOT, Utrecht.
- da Silva Facundes, S., de Freitas, M. F. P., and de Lima-Padovani, B. F. S. (2021). Number expression in Apurinã (Arawák). In Hämäläinen, M., Partanen, N., and Alnajjar, K., editors, *Multilingual Facilitation*, pages 31–42. University of Helsinki Library, Helsinki.
- de Alencar, L. F. (2021). Uma gramática computacional de um fragmento do nheengatu / A computational grammar for a fragment of nheengatu. *Revista de Estudos da Linguagem*, 29(3):1717–1777.
- de Almeida Navarro, E. (2016). *Curso de Língua Geral (nheengatu ou tupi moderno): A língua das origens da civilização amazônica*. Centro Angel Rama da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo, second edition.
- de Magalhães, J. V. C. (1876). *O selvagem*. Typographia da Reforma, Rio de Janeiro.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Eberhard, D. M., Simons, G. F., and Fennig, C. D., editors (2023). *Ethnologue: Languages of the World*. SIL International, Dallas, twenty-sixth edition.
- Freire, J. R. B. (2011). *Rio Babel: A história das línguas na Amazônia*. EdUERJ, Rio de Janeiro, second edition.
- Galves, C., Sandalo, F., Sena, T. A. d., and Veronesi, L. (2017). Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, 59(3):631–648.
- Gerardi, F. F., Reichert, S., and Aragon, C. C. (2021). TuLeD (tupían lexical database): introducing a database of a South American language family. *Language Resources and Evaluation*, 55(4):997–1015.

- Martín Rodríguez, L., Merzhevich, T., Silva, W., Tresoldi, T., Aragon, C., and Gerardi, F. F. (2022). Tupían language resources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58, Marseille, France. European Language Resources Association.
- Moore, D. (2014). Historical development of Nheengatu (Língua Geral Amazônica). In Mufwene, S. S., editor, *Iberian Imperialism and Language Evolution in Latin America*, pages 108–142. University of Chicago Press, Chicago.
- Moore, D., Facundes, S., and Pires, N. (1994). Nheengatu (Língua Geral Amazônica), its history, and the effects of language contact. In *Proceedings of the Meeting of the Society for the Study of the Indigenous languages of the Americas, July 2-4, 1993 and the Hoka-Penutian workshop, July 3, 1993*, Report / Survey of California and other Indian Languages ; 8, pages 93–118, Berkeley, CA. [University of California].
- Navarro, E., Ávila, M., and Trevisan, R. (2017). O nheengatu, entre a vida e a morte: A tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, 6(2):9–29.
- 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. European Language Resources Association (ELRA).
- Nivre, J. and Fang, C.-T. (2017). Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.
- Rodrigues, A. D. (1996). As línguas gerais sul-americanas. *Papia*, 4(2):6–18.
- Rodrigues, A. D. and Cabral, A. S. A. C. (2011). A contribution to the linguistic history of the língua geral amazônica. *ALFA: Revista de Linguística*, 55(2).
- 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, Portorož, Slovenia. European Language Resources Association (ELRA).
- Simons, G. F., Thomas, A. L. L., and White, C. K. K. (2022). Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on 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, Brussels, Belgium. Association for Computational Linguistics.
- 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'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Albertina in Action: An Investigation of its Abilities in Aspect Extraction, Hate Speech Detection, Irony Detection, and Question-Answering

Júlia da Rocha Junqueira¹, Claudio Luis Junior¹, Félix Leonel V. Silva¹
Ulisses Brisolara Côrrea¹, Larissa A. de Freitas¹

¹Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas (UFPel)
Pelotas – RS – Brazil

{julia.rjunqueira, clsmachado, flvdsilva, ulisses, larissa}@inf.ufpel.edu.br

Abstract. *The field of natural language processing has witnessed significant advances in recent decades, driven by the application of deep learning. Combined with using a neural architecture named Transformers, the advances are superior and outstanding. In this work, we used a BERT based model for the Brazilian Portuguese language, called Albertina, to tasks of Aspect Extraction, Hate Speech Detection, Irony Detection, and Question-Answering. Lastly, we compare the results in each task obtained with the BERTimbau and Albertina base and large models.*

Resumo. *O campo de processamento de linguagem natural testemunhou avanços significativos nas últimas décadas, impulsionados pela aplicação de aprendizado profundo. Combinando com o uso de uma arquitetura neural chamada Transformers, os avanços são ainda mais superiores e marcantes. Neste trabalho, usamos um modelo baseado em BERT para a língua portuguesa do Brasil, chamado Albertina, nas tarefas de Extração de Aspecto, Detecção de Discurso de Ódio, Detecção de Ironia e Perguntas-Respostas. Por fim, comparamos os resultados obtidos em cada tarefa com os modelos de base e grande de BERTimbau e Albertina.*

1. Introduction

In the last decade, the field of Natural Language Processing (NLP) has witnessed significant advances due to its application of Deep Learning (DL). The application of DL in NLP has produced more efficient and precise results for various tasks in Natural Language Understanding, particularly topic classification, sentiment analysis, question answering, and language translation [LeCun et al. 2015].

As an example of this improvement, using Transformers for NLP problems, such as BERT (Bidirectional Encoder Representations for Transformers), has become very common. Its popularity has increased because this model could represent the types of syntactic and semantic abstractions traditionally necessary for language processing. Moreover, they can model complex interactions between different levels of hierarchical information [Tenney et al. 2019].

Given this context, the development of AlbertinaPT-* [Rodrigues et al. 2023], a model based on DeBERTa, brings significant benefits for text processing in Portuguese,

allowing a more accurate and comprehensive analysis of emotions, opinions, aspects, negative speeches, and ironies present in texts written, enabling the advance of research and innovation in language technology for European Portuguese and Brazilian Portuguese.

Therefore, this work explores the tasks of Aspect Extraction (AE), Hate Speech Detection (HS), Irony Detection (ID), and Question-Answering (QA) using methods based on the BERT based model, Albertina PT-*. For this, we do our experiments in Brazilian Portuguese through the use of Albertina PT-BR, observing how the power of Albertina PT-* can be harnessed to improve the quality and accuracy of these tasks.

The paper is structured as follows: **Theoretical Background** covers the technical information relevant to understand the addressed tasks; **Related Works** reviews relevant works previously published in the literature, with a particular focus on studies covering NLP models concerning the Portuguese language; **Methodology** describes the steps taken to perform the experiments, including information about datasets, fine-tuning, and the data flux across tasks; **Experiments** shows the configuration and hyperparameters used to approach each task; **Final Remarks** summarizes the work and briefly discusses potential future studies.

2. Theoretical Background

Sentiment Analysis involves categorizing texts into positive, negative, or neutral polarities. In terms of data processing approaches, the literature describes different levels of granularity: the document-level analysis considers the entire text as a whole; the aspect-level requires additional steps, as it involves Aspect Extraction (AE) before classifying the sentiments associated with each aspect [Hoang et al. 2019]. The AE task focuses on identifying and extracting specific aspects or features discussed in a given text, usually a review. For instance, a hotel review could include specific aspects as room cleanliness, food quality, and staff friendliness [Liu 2015].

The HS task involves identifying whether various forms of communication, such as text, audio, and others, contain expressions of hatred or incite violence towards individuals or specific groups. A significant arena for spreading HS online is social media. The social media posts include paralinguistic signals (e.g., emoticons and hashtags), and their linguistic content contains plenty of poorly written text which are difficult to analyze. Another area for improvement is the lack of consensus on what constitutes HS, which makes the task difficult even for humans [Kovács et al. 2021].

The ID task corresponds to the ability to classify texts binarily, whether their respective content contains ironic behavior utilizing algorithms and models that can detect said behavior. It is difficult to determine what consists of irony and where the line is drawn when compared to sarcasm; for example, the base consensus is that irony revolves around the understanding that it is used to express the opposite of the literal meaning of what trying to be expressed by whoever is communicating, and while sarcasm can be understood the same way, it is used as a form of verbal irony, in which it carries a mocking or contemptuous tone, usually meant to mock, provoke or criticize something [Lee and Katz 1998]. Still, irony in itself can also depend on the level of aggressiveness shown and vocal clues [Van Hee et al. 2018].

The QA task combines several research fields, such as Information Retrieval, Information Extraction, and NLP. The methods used for the task aim to solve and

propose answers relevant to the question selected. The task can be divided into three modules: question classification, information retrieval, and answering extraction [Allam and Haggag 2012]. Question classification returns the type of answer that the question informed is expecting; for example, if you ask “What year the computer was invented?” the model is expected to return a year as an answer. Information retrieval returns search results based on the question submitted and its type; if it does not find information that matches, no further processing is carried out. Answering extraction returns the answer to the question asked.

3. Related Works

Below we describe some works in the literature about AE, HS, ID, and QA. Furthermore, we describe the work proposed by [Rodrigues et al. 2023], the Albertina model’s.

In 2022, [da Silva et al. 2022] proposed the first shared task dedicated to identifying aspects and extracting the polarity in texts written in Portuguese, the Aspect-Based Sentiment Analysis in Portuguese (ABSAPT). ABSAPT comprised two sub-tasks: Aspect Extraction (AE) and Aspect Sentiment Classification (ASC). The results showed that with the ABSAPT 2022 hotel reviews dataset, the BERT methods, specifically “base-multilingual-cased” and “base-portuguese-cased”, to AE, achieved an Accuracy of 0.67 [Gomes et al. 2022]. These studies have showed promising outcomes in enhancing the accuracy of Sentiment Analysis for Portuguese texts.

The solution to HS task proposed by [Leite et al. 2020] split ToLD-BR dataset into three parts: 80% for training, 10% for development, and 10% for testing. They utilized Bag-of-Words (BoW) to represent the examples and an AutoML model to establish the baseline model (BoW + AutoML). To accomplish this, they employed the auto-sklearn library for BERT-based models. The simple transformers library was utilized for convenient training and evaluation. Default arguments were used for parameter tuning, and a specific seed was defined to ensure reproducibility. Two versions of BERT, namely mBERT and BERTimbau base [Souza et al. 2020], were employed. The resulting F-Measure scores were 0.74 for BoW + AutoML, 0.75 for mBERT, and 0.76 for BERTimbau base.

In 2021, [Corrêa et al. 2021] proposed the first shared task dedicated to identifying the presence of irony in texts (tweets and news) written in Portuguese. The results showed that, with the IDPT 2021 tweets dataset, the classical feature-based models outperformed Deep Learning methods, achieving a BAcc of 0.52. [Jiang et al. 2021] introduced a solution to address the problem by utilizing BERTimbau, weight loss, and ensemble learning. The author claimed that the best-performing strategy involved leveraging two datasets used in IDPT 2021 for assisting in model classification and generalization, and this strategy achieved a BAcc of 0.48. Due to the relatively small size of the IDPT 2021 dataset [Subies 2021] they opted to employ Data Augmentation techniques. [Jiang et al. 2021] applied random masking to 15% of the tokens and utilized BERTimbau base and hyperparameter Grid Search to predict the masked tokens. The experiments with BERTimbau presented a BAcc of 0.49.

The QA task development was based on Guillou’s work, using the BERTimbau model, fine-tuned on the SQUAD v1.1 in the Portuguese dataset [Guillou 2021]. [Spindola et al. 2021] also cites Guillou’s work in their paper “Portuguese-Based Ques-

tion Answering System about the Blue Amazon”. The authors combined BERTimbau model fine-tuned by Guilou with BertViz to visualize the attention weights and compared the results on their dataset Blue Amazon QA [Spindola et al. 2021]. Guillou, in his work, has split the dataset into two parts, training and validation. Since the dataset contains many words in each paragraph, the length was limited to 384 characters, allowing only one long example in the dataset, to give it different input features. Then, the Trainer API for feature-complete training in PyTorch was used to fine-tune and evaluate the model. The results obtained were 70.49% of Exact Match and 0.82 of F-Measure, using different hyperparameters on the BERTimbau base model.

According to [Rodrigues et al. 2023], the Albertina model’s ability to achieve superior performance with less training time/computation likely results from resorting to all pre-trained layers, including the first layer, concerning word embeddings and the last layer, concerning masked token prediction, in contrast to the common practice in the literature of resetting these two layers to random weights to continue the pre-training. The tasks the author chose to demonstrate the model’s functioning were: Remote Procedure Call, Semantic Textual Similarity, Recognizing Textual Entailment, and WNLI. In our experiments, we used the Albertina model’s in AE, HS, ID, and QA task.

4. Methodology

Our work is composed of four main steps (Figure 1). Initially, the Albertina model is used. After, we applied fine-tuning in the AE, HS, ID, and QA tasks. And we test in datasets ABSAPT 2022 [da Silva et al. 2022], ToldBR [Leite et al. 2020], IDPT 2021 [Corrêa et al. 2021], and SQUAD v1.1 [Rajpurkar et al. 2016]. Finally, we analyzed the results obtained in each task, comparing them to BERTimbau.

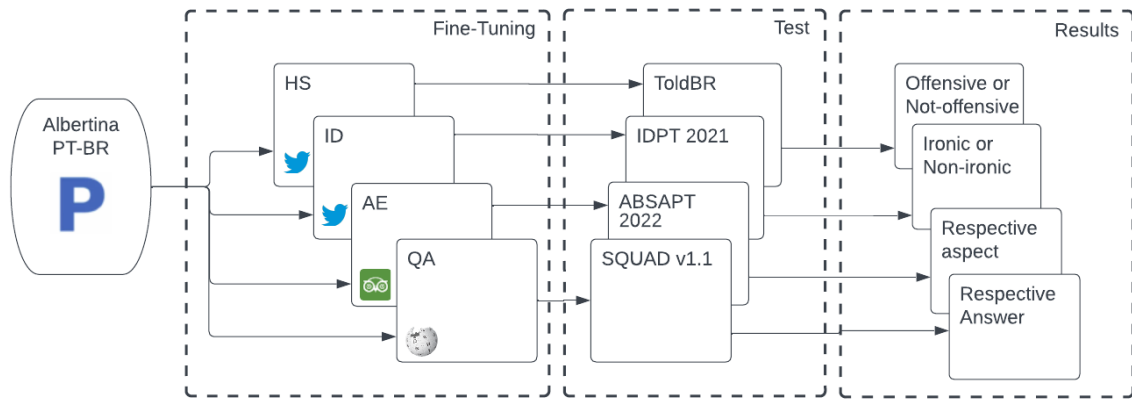


Figure 1. Methodology of this work.

For the AE task, we used the Trainer API¹ from Huggingface for the training and fine-tuning, and lastly, the Evaluate library for validating and evaluation of the results. The dataset utilized for AE task comprises reviews sourced from TripAdvisor, which were compiled by [da Silva et al. 2022]. The training data consists of 847 reviews, divided in 77 aspects, and includes 3111 sentiment polarity annotations. Among these annotations are 2112 positive examples, 472 neutral examples, and 527 negative examples. On the

¹https://huggingface.co/docs/transformers/main_classes/trainer

other hand, the test dataset comprises 184 reviews involving 70 aspects and 686 sentiment polarity annotations. Of these annotations, 450 are positive, 105 are neutral, and 131 are negative.

For the HS task, the ToLD-BR dataset was used. It comprises tweets gathered between July and August 2019 utilizing GATE Cloud’s Twitter Collector tool². Two distinct strategies were employed for tweet collection. The first strategy involved searching for specific keywords and predefined hashtags like “gay”, “little woman”, and “northeasterner”. The second strategy involved gathering tweets that mentioned influential figures such as Brazil’s former President Jair Bolsonaro, and soccer player Neymar Jr. This method imposed no restrictions on keywords or hashtags, resulting in the collection of over 10 million unique tweets, out of which 21000 were randomly selected to compose the dataset. It should be noted that the first strategy accounted for 60% of the collected data. To annotate the dataset, 42 annotators were involved in classifying 1500 tweets as LGBTQ+phobia, obscene, insult, racism, misogyny, or xenophobia. Ultimately, the dataset consisted of 9245 offensive tweets and 11693 non-offensive tweets, each classified by three annotators.

For the ID task, the dataset utilized was the IDPT 2021 tweets [Corrêa et al. 2021], and was manually classified by linguistics and computer science students. The dataset is divided into two columns, one composed of text, the actual tweets, and another that determines whether or not what is written in the said tweet is a form of irony/sarcasm. We differentiate what composed ironic and non-ironic as follows: (1) Ironic are sentences contradict the meaning between what is intended and what is written, e.g.: “*Que time horrível esse do Vasco, quase fez gol!!!*” [“*What a horrible team Vasco is, (they) almost scored!!!*”]; (2) Non-ironic are sentences that do not contain linguistic mechanisms that alternate their meaning, e.g.: “*Frustração hoje tem nome: Economia! — se sentindo triste*” [“*Frustration today has a name: Economy! — feeling sad*”]. Tweets were composed of 12736 ironic, and 2476 non-ironic. The testing dataset has a similar constitution but is a lot smaller, consisting of only 300 tweets, 177 tweets ironic and 123 tweets non-ironic.

For the QA task, the Trainer API for feature-complete training in PyTorch was used to fine-tune and evaluate the results, such as [Guillou 2021]. The dataset used was the SQUAD v1.1-PT, created by automatically translating the content of SQUAD using the Google Cloud API. The dataset was split into 87599 rows of paragraphs for train and 10570 for validation. Its data is composed of a title, a context, a question, and an answer, based on Wikipédia articles, where the answer to every question is a segment of text from the corresponding reading context [Rajpurkar et al. 2016].

5. Experiments

For most experiments, we used a batch size of 8, 3 epochs, a learning rate of $1 * 10^{-5}$, loss function CrossEntropy and optimizer AdamW (Table 1). The model training process is tailored to balance computational limitations with the need to achieve reasonable model performance. It is important to note that these hyperparameter choices were made based on constraints and may not necessarily represent the optimal configuration for the problem.

²<https://cloud.gate.ac.uk/info/help/twitter-collector.html>

Table 1. Hyperparameters Across Experiments.

Hyperparameters	Base	Large
Attention Heads	12	16
Batch Size (*)	8	2
Epochs	3	3
Hidden Size	768	1536
Hidden Layers	12	24
Learning Rate	1e-5	1e-5
Loss Function	CrossEntropy	CrossEntropy
Optimizer	AdamW	AdamW
Parameters	100 M	900 M

*: The QA experiment is an exception regarding the displayed batch size. We use a batch size of 16 and 8 for the base and large models, respectively.

The model Albertina (PT-BR) was tested on four tasks AE, ID, HS, and QA. Each test dataset was evaluated on several metrics, such as Accuracy (Acc), Precision, Recall, and F-Measure [Brownlee 2016], except for the QA task, which was evaluated based on Exact Match (EM) and F-Measure only.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Instances} \quad (1)$$

$$Precision = \frac{True\ Positives}{True\ Positives + True\ Negatives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

$$F - Measure = 2 \frac{Precision.Recall}{Precision + Recall} \quad (4)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (5)$$

$$ExactMatch = \frac{TruePositives}{TotalNumberofInstances} * 100 \quad (6)$$

In Table 2, we show the results of our experiments for each task with the models BERTimbau base and Albertina PTBR base. In Table 3, we present the results of our experiments for each task with the models BERTimbau large and Albertina PTBR large.

To ensure the selection of meaningful aspects, we employ a filtering process in the AE task, excluding aspects with fewer than 20 occurrences, only the remaining 25 aspects for the experiments. It is possible to observe that the results from both models were low compared to the other tasks and the work presented in the Related Works session, which resulted in a 67% Accuracy by [Gomes et al. 2022]. This results from our technical limitations since this task demand a higher processing power and requires a larger memory

Table 2. Results Obtained Using BERTimbau and Albertina PT-BR Base Models.

	Task	Dataset	Acc	Precision	Recall	F-Measure	EM%
BERTim.	AE	ABSAPT 2022	0.26	0.21	0.26	0.19	-
	HS	ToLD-BR	0.88	0.89	0.88	0.88	-
	ID	IDPT 2021	0.41	0.36	0.41	0.25	-
	QA	SQUAD v1-PT	-	-	-	0.56	43.29
Albertina	AE	ABSAPT 2022	0.22	0.12	0.22	0.13	-
	HS	ToLD-BR	0.78	0.72	0.77	0.74	-
	ID	IDPT 2021	0.40	0.40	0.99	0.57	-
	QA	SQUAD v1-PT	-	-	-	0.57	45.12

Table 3. Results Obtained Using BERTimbau and Albertina PT-BR Large Models.

	Task	Dataset	Acc	Precision	Recall	F-Measure	EM%
BERTim.	AE	ABSAPT 2022	0.27	0.23	0.27	0.22	-
	HS	ToLD-BR	0.89	0.90	0.89	0.89	-
	ID	IDPT 2021	0.40	0.16	0.40	0.22	-
	QA	SQUAD v1-PT	-	-	-	0.62	47.15
Albertina	AE	ABSAPT 2022	0.21	0.04	0.20	0.07	-
	HS	ToLD-BR	0.58	0.34	0.58	0.43	-
	ID	IDPT 2021	0.41	0.41	1.0	0.58	-
	QA	SQUAD v1-PT	-	-	-	0.32	47.30

capacity. Furthermore, BERTimbau returned better results than Albertina, in both models, base and large.

For HS, the results indicated that the Albertina PT-BR model was inferior in every aspect compared to the results using BERTimbau. The model achieved an Accuracy score of 78% (base) and 58% (large), while the BERTimbau model performed significantly better with an Accuracy of 88% (base) and 89% (large). Similarly, Recall, was significantly higher for the BERTimbau model with a score of 88% (base) and 89% (large). On the other hand, Albertina PT-BR achieved a Recall score of 77% (base) and 58% (large). Lastly, the F-Measure, further confirms the superiority of the BERTimbau model. BERTimbau achieved an F-Measure of 88% (base) and 89% (large), while Albertina PT-BR only managed a score of 74% (base) and 43% (large). The higher F-Measure of BERTimbau indicates a better balance between Precision and Recall, reflecting its overall better performance in identifying HS instances accurately. A significant difficulty in detecting hate speech is identifying the context of what is or is not hate speech, many of the texts are also poorly written, such as “*Ui Nooooooofa que lindo fofa Nosso galao e mara ne amiga rajkazblanks*” [“*Wow how beautiful cute Our gallon and mara right friend rajkazblanks*”], another factor that affected the results achieved were the hyperparameters used, which were reduced by the limitation of machine resources during the experiments performed in this work.

The model was loaded and fine-tuned for ID, and with the Trainer API, the metric

was evaluated. The metrics used for the evaluation were Accuracy, Precision, Recall, and F-Measure, and that, along with the Actual Values and Predicted Values, formed True Positives, True Negatives, False Positives, and False Negatives. This was possible because the task was a binary evaluation with only two feasible outcomes. In the base model, compared to the BERTimbau utilized in [Corrêa et al. 2021], the result for Accuracy in the Albertina PT-BR was 40.67%, compared to 41%. The Precision results marked 40.8% compared to the 36%. The Recall results marked 99.18% compared to the 41%. The F-Measure results were 57.81% compared to 25%. The large models showed the following results: Accuracy 41%, compared to 40%. The Precision results marked 41% compared to the 16%. The Recall results marked 100%, compared to the 40%. The F-Measure results were 58.15% compared to 22%. These general results show that the Albertina PT-BR model does show an increase in all aspects, which can be compared to previous results with BERTimbau. ID can be a challenging task to obtain due to the complex and context-dependent nature of the language, and without context being provided to the model, it is likely to miss some tweets such as “*O problema do Brasil não é a violência, são as vítimas.*” [“*The problem with Brazil is not violence, but the victims.*”], and that, combined with the fact that lower parameters were used in order to achieve results because of hardware limitations, result percentages were worsened.

For the QA task, after the fine-tuning of the model, we used the Trainer API to evaluate the metrics. In this task, only two metrics were used for the evaluation, F-Measure, and Exact Match, since the task’s resolution is composed of answers, not a binary representation. We obtained 45.12% of EM on Albertina PT-BR, compared to 43.29% to BERTimbau, both on the base versions. In the large versions, we obtained 47.30% of EM on Albertina PT-BR, in comparison 47.15% to BERTimbau. It is observable that our results cannot be compared by EM to the other tasks since it evolves on another kind of representation, but, compared to the work of [Guillou 2021], it was considerably lower side-by-side to his results, 70.49%. This is due to our hyperparameters choice of configuration; we precisely choose to lower our hyperparameters so that all the tasks could run without achieving technical limitations. In addition, the EM metric tends to be very limited since it needs to account for minor variations, such as differences in punctuation, capitalization, or word order. Therefore, if a predicted answer is semantically correct but slightly different than the correct match, it is considered wrong.

6. Final Remarks

With this work, we can conclude that the Arbertina PT-BR model can be better in some tasks when compared to BERTimbau model, such as in ID and QA. In other tasks, such as HS, the results can be inferior in every aspect, including Accuracy, Precision, Recall, and F-Measure, significantly reducing the score percentage.

Regarding our future research, there are several aspects that can be considered for further investigation. Firstly, the scope can be expanded by integrating another datasets, enabling a more comprehensive analysis. Additionally, it would be beneficial to evaluate the models in relation to different tasks and implement preprocessing techniques to address any data imbalance issues, thus enhancing the accuracy of the datasets. Moreover, it is worthwhile to explore alternative hyperparameters for tasks that exhibit lower performance, such as AE and HS, in order to potentially achieve improved results.

References

- Allam, A. M. N. and Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).
- Brownlee, J. (2016). *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*. Machine Learning Mastery.
- Corrêa, U. B., Coelho, L., Santos, L., and de Freitas, L. A. (2021). Overview of the idpt task on irony detection in portuguese at iberlef 2021. *Procesamiento del Lenguaje Natural*, 67.
- da Silva, F. L. V., da S. Xavier, G., Mensenburg, H. M., Rodrigues, R. F., dos Santos, L. P., Araújo, R. M., Corrêa, U. B., and de Freitas, L. A. (2022). Absapt 2022 at iberlef: Overview of the task on aspect-based sentiment analysis in portuguese. *Procesamiento del Lenguaje Natural*, 69.
- Gomes, J. R. S., Garcia, E. A. S., Junior, A. F. B., Rodrigues, R. C., Silva, D. F. C., Maia, D. F., da Silva, N. F. F., Filho, A. R. G., and da Silva Soares, A. (2022). Deep learning brasil at ABSAPT 2022: Portuguese transformer ensemble approaches. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2022), co-located with the 38th Conference of the Spanish Society for Natural Language Processing (SEPLN 2022)*, Online. CEUR. org, Online. CEUR. org.
- Guillou, P. (2021). Portuguese bert base cased qa (question answering), finetuned on squad v1.1.
- Hoang, M., Bihorac, O. A., and Rouces, J. (2019). Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.
- Jiang, S., Chen, C., Lin, N., Chen, Z., and Chen, J. (2021). Irony detection in the portuguese language using bert. *Proceedings* <http://ceur-ws.org> ISSN, 1613.
- Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media. *SN Computer Science*, 2.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.
- Lee, C. J. and Katz, A. N. (1998). The differential role of ridicule in sarcasm and irony. *Metaphor and symbol*, 13(1):1–15.
- Leite, J. A., Silva, D. F., 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*.
- Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

- 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. *arXiv preprint arXiv:2305.06721*.
- 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*.
- Spindola, S., José, M. M., Oliveira, A. S., Cação, F. N., and Cozman, F. G. (2021). Interpretability of attention mechanisms in a portuguese-based question answering system about the blue amazon. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 775–786. SBC.
- Subies, G. G. (2021). Guillemgsubies at idpt2021: Identifying irony in portuguese with bert. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2021), co-located with the 37th Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Online. CEUR. org*, pages 910–916, Online. CEUR. org.
- Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.
- Van Hee, C., Lefever, E., and Hoste, V. (2018). SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Studying the Dependence of Embedding Representations on the Target of NLP Tasks

Bárbara Stéphanie Neves Oliveira¹, Ticiana L. Coelho da Silva¹, José A. F. de Macêdo¹

¹Insight Data Science Lab

Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brazil

{barbaraneves,ticianalc,jose.macedo}@insightlab.ufc.br

Abstract. *In many human languages, linguistic units represent text structure. Vector semantics is used in NLP to represent these units, known as embeddings. Evaluating the learned representations is crucial for identifying critical differences between the diverse existing embedding models in task-specific selection. However, the evaluation process is complex, with two approaches: intrinsic and extrinsic. While useful, aggregated evaluations often lack consistency due to result misalignment. This work investigates the dependencies and correlations between embeddings and NLP tasks. The goal is how to initially verify if the embeddings' dimensions (i.e., features) depend on the final task. The study then explores two research questions and presents findings from experiments.*

1. Introduction

In many human languages, most information about the structure of texts can be represented in the form of linguistic units. Understanding how to learn textual representations using Deep Learning techniques is a crucial area of research in Natural Language Processing (NLP) [Jurafsky and Martin 2018, Oliveira et al. 2022]. This focus has given rise to various architectures that aim to model words or other linguistic units such as characters, sentences, or documents.

Embeddings are a fundamental concept in NLP that serve as a form of textual representation. They are numerical vectors that encode both the meaning and contextual information of linguistic units within a given language. Many methods have been developed to generate embeddings, from more straightforward approaches to sophisticated techniques [Torregrossa et al. 2021, Oliveira et al. 2022].

The field of NLP has yet to converge on a universal embedding method and scale it sufficiently to provide state-of-the-art results on all tasks [Ignat et al. 2023, Muennighoff et al. 2022]. Consequently, constructing effective NLP pipelines with high-quality input representations remains challenging, especially with abundantly available techniques. This leads to confusion about which model provides practitioners with the best performance for their embedding use case. Thus, assessing the learned representations is vital in identifying the critical distinctions between various embedding models, enabling the selection of the most suitable one for a specific task [Boggust et al. 2022, Bakarov 2018].

Evaluating embeddings involves two primary approaches: extrinsic and intrinsic. While extrinsic evaluation guarantees practical performance, intrinsic evaluation offers insights into the inherent quality of embeddings [Jurafsky and Martin 2018]. However, further advancements are needed to enhance the evaluation and comparison process, bridging

existing gaps and unlocking the full potential of these powerful language representation models [Schnabel et al. 2015, Bakarov 2018, Torregrossa et al. 2021].

In recent years, researchers have recognized the importance of addressing the preliminary verification of embeddings before utilizing them to represent a corpus or corpora in an NLP task [Boggust et al. 2022, Muennighoff et al. 2022]. This research then tackles the fundamental challenge of measuring (i.e., with heuristics or numerical measures) the dependencies and correlations between the input textual representations and the ultimate objective of an NLP task.

Contributions. Overall, the main contributions of this paper can be summarized as follows:

- Investigate the crucial step of examining whether the learned vectors, used as features, are relevant to the final task, ensuring high-quality representations.
- Present two main research questions to guide the study, providing detailed discussions, experimental setups, and results for each one.
- Conduct extensive experiments exploring whether numerical measures can determine the dependence between input embeddings and their suitability for a specific NLP task.

2. Related Work

To address the challenge of verifying input representations for NLP tasks, research works mainly focuses on three topics, presented below. This work addresses the first topic, while the remaining topics are summarized to provide an overview of the current state.

Explainability and interpretability. General techniques use different tools to understand model predictions, feature importance, and decision-making processes [Hamilton et al. 2016, Ribeiro et al. 2016, Shrikumar et al. 2017, Carter et al. 2019]. Unlike other methods, this work compares embeddings learned by different models using a global measure, considering that internal representations can vary.

Visual embedding techniques and tools. To reason about and interpret the learned representations, [Heimerl and Gleicher 2018], [Liu et al. 2019b], and [Boggust et al. 2022] propose interactive or static systems for exploring embeddings via direct projection manipulation, interactively filtering, and reconfiguring visual forms.

Methods for comparing embedding spaces. To compare vector spaces, research works perform alignment through linear transformation [Chen et al. 2018] or nearest neighbors and co-occurrences over time [Heimerl and Gleicher 2018, Wang et al. 2018b], relationship analysis between node metrics and graph embeddings [Li et al. 2018], and evaluation of vector consistency across latent embedding spaces [Liu et al. 2019b, Boggust et al. 2022].

3. Research Questions and Discussion

Before delving into the experiments, the main research questions that guide this study are:

RQ1 Can heuristics or numerical measures determine the dependence between the input embeddings and their suitability for a particular NLP task?

Context. Some model performances can degrade when including input features irrelevant to the target labels. Typically, feature selection methods are intended to reduce the number of input features to those considered most beneficial based on statistical tests [Butcher and Smith 2020].

Quantitative investigation. This research question aims to analyze the quality of pre-trained input embeddings¹ by applying an existing feature selection measure to different types of representations and corpora. The purpose is not to select the best dimensions for a specific task but to identify which embedding approach has more dimensions with high scores indicating strong dependence between input and output. The experiments focus on embedding representations for the training set.

RQ2 To what extent does the model developed to solve an NLP task affect the transferability of the input embeddings?

Context. After using a heuristic or numerical measure to assess the suitability of input embeddings for an NLP task, the question arises: what happens when the same method is employed now to evaluate the quality of the embeddings alongside the predicted labels generated by the model?

Qualitative investigation. This research question explores the relationship between the architecture of an NLP task, the linguistic knowledge encoded in pre-trained input embeddings, and their transferability. An extrinsic evaluation approach is employed to investigate this, where a model is trained using different representations. The objective is to determine whether models with high evaluation metrics also exhibit high dependency values. In this context, the feature selection measure is implemented on the test set, considering both the actual labels and the predicted ones produced by the model.

Statements. Additionally, since different embeddings ends-up producing similar results for the same model [Muennighoff et al. 2022], the following scenarios are also considered: (i) if the feature selection measure indicates high dependency values, yet the model still performs poorly, the issue may lie with the remaining network components of the model; and (ii), likewise, if the measure suggests low dependency values, but the model achieves high-quality results, subsequent layers beyond the textual representation may impact transferability. The ultimate goal of this research question is to determine the interchangeability of the mentioned statements.

4. Experimental Setup

4.1. Probing Task

The concept of probing tasks introduced by [Shi et al. 2016] and [Adi et al. 2016] involves using a pre-trained encoder (e.g., embeddings) to train a classifier or decoder that focuses on simple linguistic properties of sentences [Conneau et al. 2018]. If the classifier succeeds, the pre-trained encoder representations contain sufficient information to solve the task effectively. Given the distinct aspects of the research questions, the experimentation focused only on the Text Classification task as a probing task, specifically Sentiment Analysis².

¹This work focuses on pre-trained and fine-tuned embeddings, which have become a trend in NLP systems and a key component of state-of-the-art models [Liu et al. 2019a].

²It is worth noting that this task belongs to the group of tasks used as an extrinsic evaluation method.

Table 1 summarizes the main statistics of the datasets used in this paper: one of them, CoLA [Warstadt et al. 2019], is part of the benchmark GLUE [Wang et al. 2018a], and the remaining three, IMDb [Maas et al. 2011], SST-2 [Socher et al. 2013], and Sentiment140 [Go et al. 2009], are generic datasets widely utilized for Text Classification/Sentiment Analysis task.

Sentiment Classification Dataset	# Corpus	# Class	Is Balanced?	# Per Class		Language
				0	1	
IMDb [Maas et al. 2011]	50,000	2	Yes	25,000	25,000	English
SST-2 [Socher et al. 2013]	68,219	2	No	30,207	38,012	
CoLA [Warstadt et al. 2019]	9,594	2	No	2,850	6,744	
Sentiment140 [Go et al. 2009]	160,000	2	Yes	79,849	80,151	

Table 1. Statistics of corpora used in experiments. To ensure comparability with other datasets, 10% of the Sentiment140 training data was randomly selected to maintain comparable text amounts. The classification values in the table correspond to combined subsets within each corpus, including training, test, or validation sets.

All corpora are in English and were loaded via Hugging Face³. Although Sentiment140 originally had three classes, only the available training set with two classes was used. Also, hold-out validation was conducted by combining all properly annotated texts from each corpus. The data was split into 70% training and 30% test sets.

4.2. Feature Selection Measure

Mutual Information (MI) is a statistical measure that quantifies the mutual dependence or information shared between two variables [Fano 1961]. In the context of NLP and embeddings, MI can be used to assess how well the embedding representation captures relevant information about the input text and its corresponding labels in a given task. However, there are some considerations to keep in mind:

Advantages. MI captures relevant information since it measures the relevance of the embedding representation to the task (i.e., higher MI indicates more relevant information). Additionally, MI can deal with non-linear dependencies since it helps model complex relationships between text and task targets.

Challenges. The accurate estimation of MI can be difficult, especially for high-dimensional embeddings. While MI provides a quantitative measure, understanding the specific linguistic or semantic aspects captured or neglected may be challenging.

In summary, MI can be a valuable tool for assessing the quality of embedding representations for NLP tasks [Zhelezniak et al. 2020]. To tackle the listed challenges, it is crucial to complement MI with other evaluation techniques. Incorporating task performance (i.e., qualitative analysis) will be essential in the results section.

Here, MI is used with Scikit-learn⁴. The `discrete_features` parameter was modified to consider continuous features. As mentioned as one of the challenges, using MI with dense representations has been difficult since it can have some issues estimating

³<https://huggingface.co/datasets>

⁴https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html

MI for continuous random variables [Zhelezniak et al. 2020]. Yet, the aim is to assess Scikit's usability for such representations.

4.3. Probing Model and Parameters

A basic model architecture for Sentiment Analysis comprises two intermediate layers with 32 units each and a final layer for sentiment prediction with sigmoid activation. The models are trained using the Keras library, a high-level API of TensorFlow, for up to 300 epochs with early stopping and a patience of five.

4.4. Pre-Trained Embedding Models

The experimentation involves three widely used publicly available models for English embeddings: GloVe [Pennington et al. 2014] and fastText [Bojanowski et al. 2016], which are static/classic embeddings, and DistilBERT [Sanh et al. 2019], a contextual representation model that has 40% fewer parameters than the original BERT Base.

The pre-trained static embeddings GloVe⁵ and fastText⁶ were loaded via the Flair library for NLP⁷, having 300-dimensional vectors each. The DistilBERT pre-trained model⁸ was instantiated using the Transformers library from Hugging Face. By default, the hidden states of all Transformer-based model layers are concatenated to produce the embeddings, generating vectors with 768 dimensions.

Sentence embeddings. To accommodate the requirements of the Scikit-learn MI function, the pre-trained representations were employed as sentence embeddings. The pooling operation used for static and contextual embeddings gives the mean of all words in the sentence. The texts with the pre-trained static sentence embeddings were embedded via Flair. The sentence embedding matrices were extracted after training the models to be used as input to the MI function.

5. Experimental Results

5.1. RQ1 Results

Table 2 contains the performance of the models during training via accuracy results and some info about MI scores (maximum and mean values). Table 3 retains precise info about the MI scores distribution, showing descriptive statistics of the percentiles. The following observations can be made in greater detail:

Pre-trained sentence embedding representation	IMDb			SST-2			CoLA			Sentiment140		
	Training Acc		MI Scores	Training Acc		MI Scores	Training Acc		MI Scores	Training Acc		MI Scores
	Best Epoch	Max	Mean	Best Epoch	Max	Mean	Best Epoch	Max	Mean	Best Epoch	Max	Mean
GloVe (300d)	0.879	0.052	0.011	0.950	0.067	0.037	0.923	0.012	0.012	0.792	0.032	0.011
fastText (300d)	0.881	0.070	0.010	0.963	0.065	0.035	0.874	0.014	0.002	0.817	0.040	0.011
Fine-tuned GloVe (300d)	0.985	0.556	0.306	0.984	0.279	0.097	0.958	0.060	0.014	0.953	0.155	0.027
Fine-tuned fastText (300d)	0.987	0.602	0.384	0.980	0.300	0.117	0.940	0.143	0.033	0.929	0.220	0.050
DistilBERT (768d)	0.908	0.089	0.013	0.918	0.117	0.040	0.911	0.022	0.003	0.794	0.040	0.008

Table 2. Model performance results during training and MI scores key information. The best results for the accuracy metric are highlighted in bold.

⁵<https://nlp.stanford.edu/projects/glove>

⁶<https://fasttext.cc/docs/en/english-vectors.html>

⁷<https://flairnlp.github.io/docs/intro>

⁸https://huggingface.co/docs/transformers/model_doc/distilbert

Pre-trained sentence embedding representation	IMDb			SST-2			CoLA			Sentiment140		
	MI Scores Percentiles											
	25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
GloVe (300d)	0.005	0.008	0.015	0.329	0.036	0.040	0.000	0.000	0.003	0.009	0.010	0.013
fastText (300d)	0.004	0.007	0.012	0.030	0.033	0.037	0.000	0.000	0.003	0.008	0.009	0.013
Fine-tuned GloVe (300d)	0.207	0.323	0.434	0.049	0.074	0.128	0.003	0.012	0.022	0.010	0.019	0.036
Fine-tuned fastText (300d)	0.282	0.462	0.552	0.062	0.098	0.171	0.011	0.025	0.051	0.016	0.036	0.069
DistilBERT (768d)	0.005	0.009	0.018	0.036	0.037	0.044	0.000	0.001	0.005	0.004	0.006	0.010

Table 3. Descriptive statistics with the percentiles of the MI scores distribution for the sentence embeddings of each training set.

Fine-tuning process. The static sentence embeddings, such as those from GloVe and fastText, were fine-tuned during training. The results include the frozen and fine-tuned versions. Although this deviates from the standard approach of maintaining encoder architecture agnosticism in probing tasks, it allows an understanding of the extent of the dependency introduced by fine-tuning these static vectors. On the other hand, the fine-tuning of DistilBERT was not performed to enable a more focused investigation of its original internal layers and their contributions to the task.

High MI score values for fine-tuned representations. Indeed, as reported in Tables 2 and 3, fastText’s updated embeddings showed the highest dependency across all datasets, closely followed by fine-tuned GloVe embeddings. On the other hand, the non-updated sentence embeddings generally had MI scores much closer to 0. Although DistilBERT has many dimensions with score values close to 0, it exhibits a more comprehensive range of scores among the non-tuned representations.

Good performance of probing models during training. Another observation is that most models converged well during training, obtaining accuracies close to or greater than 0.90, except for GloVe and DistilBERT on the Sentiment140 training data. The RQ2 will verify if the models are ideal and are on the borderline between underfitting and overfitting.

RQ1 answer. As a response to RQ1, the MI measure does not readily indicate which sentence representations are sufficient to solve the different Sentiment Analysis tasks. Despite the MI measure lacking evidence, these results are still valuable as part of the research, which includes attempts beyond initial expectations.

Updating sentence embeddings during training logically improves the correlation between input embeddings and the task objective. Also, note that the MI values are low for CoLA and Sentiment140. This will be discussed further in the next section.

5.2. RQ2 Results

The objective of the RQ2 is to observe the semantic transferability of embeddings with the MI measure and an extrinsic evaluation. The analysis conducted using the MI measure on the training set was similarly applied to the test set. Minor changes were expected in the distribution of MI scores, as the main characteristics of the data were retained for both sets. Tables 4 and 5 present descriptive statistics of the percentile distribution of MI scores, including actual and predicted classes. Further examination reveals the following detailed observations:

Conflicting MI score values for CoLA. As expected, the distributions are equivalent to

Pre-trained sentence embedding representation	IMDb			SST-2			CoLA			Sentiment140		
	MI Scores Percentiles											
	25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
GloVe (300d)	0.003	0.007	0.014	0.014	0.017	0.022	0.000	0.001	0.006	0.007	0.010	0.011
fastText (300d)	0.002	0.006	0.011	0.015	0.018	0.021	0.000	0.001	0.006	0.007	0.010	0.013
Fine-tuned GloVe (300d)	0.168	0.264	0.336	0.029	0.053	0.102	0.000	0.002	0.008	0.011	0.018	0.032
Fine-tuned fastText (300d)	0.225	0.349	0.403	0.038	0.069	0.132	0.000	0.002	0.007	0.016	0.032	0.061
DistilBERT (768d)	0.003	0.008	0.017	0.016	0.019	0.026	0.000	0.002	0.007	0.003	0.005	0.009

Table 4. Descriptive statistics with the percentiles of the MI scores distribution for the sentence embeddings of each test set with the *actual* classes.

Pre-trained sentence embedding representation	IMDb			SST-2			CoLA			Sentiment140		
	MI Scores Percentiles											
	25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
GloVe (300d)	0.003	0.008	0.015	0.016	0.019	0.023	0.003	0.008	0.014	0.012	0.014	0.017
fastText (300d)	0.003	0.008	0.014	0.015	0.018	0.023	0.005	0.011	0.017	0.012	0.015	0.022
Fine-tuned GloVe (300d)	0.184	0.283	0.384	0.031	0.056	0.106	0.006	0.014	0.022	0.015	0.025	0.039
Fine-tuned fastText (300d)	0.244	0.416	0.510	0.041	0.075	0.143	0.010	0.021	0.032	0.021	0.040	0.071
DistilBERT (768d)	0.003	0.009	0.020	0.016	0.020	0.027	0.000	0.003	0.009	0.004	0.007	0.014

Table 5. Descriptive statistics with the percentiles of the MI scores distribution for the sentence embeddings of each test set with the *predicted* classes.

those for **RQ1**. However, the CoLA dataset has the most incompatible distributions, with score values lower than those obtained for the training set. Note that for the predicted classes, CoLA’s MI scores are much more similar to those for the training set, indicating that the model can accurately maintain the variance and patterns learned during training.

MI score values indicating overfitting for CoLA and Sentiment140. Also, for the CoLA dataset, it is possible to deduce, based on the MI scores, that the model may not have performed well for the test set, which indicates the existence of overfitting. Overfitting can also be noticed for Sentiment140 by some subtle changes when comparing the distributions of the scores of the predicted classes with the actual ones. The existence or not of overfitting in the models will be further analyzed. The fact that the MI scores could reveal this unexpected model behavior is intriguing.

Extrinsic evaluation. Table 6 reports the accuracies obtained between the ground truth and what was predicted by the probing models, and some info about the MI scores. Table 7 has more extrinsic evaluation results for other metrics.

Pre-trained sentence embedding representation	IMDb			SST-2			CoLA			Sentiment140		
	Test Acc	MI Scores		Test Acc	MI Scores		Test Acc	MI Scores		Test Acc	MI Scores	
		Max	Mean		Max	Mean		Max	Mean		Max	Mean
GloVe (300d)	0.830	0.050	0.010	0.879	0.049	0.019	0.625	0.024	0.004	0.500	0.267	0.010
fastText (300d)	0.857	0.072	0.009	0.887	0.051	0.019	0.609	0.021	0.004	0.500	0.042	0.010
Fine-tuned GloVe (300d)	0.879	0.406	0.241	0.910	0.242	0.072	0.577	0.023	0.004	0.497	0.139	0.025
Fine-tuned fastText (300d)	0.880	0.430	0.288	0.912	0.253	0.088	0.615	0.024	0.004	0.499	0.190	0.044
DistilBERT (768d)	0.866	0.097	0.013	0.867	0.094	0.022	0.677	0.027	0.004	0.501	0.044	0.007

Table 6. Model performance results for the test set and MI scores key information. The best results for the accuracy evaluating metric are highlighted in bold.

Overfitting for CoLA and Sentiment140. The models with the different sentence embeddings have high and very close accuracies for the IMDb and SST-2 datasets. It is possible to state through these results that the CoLA and Sentiment140 models suffered overfitting.

Pre-trained sentence embedding representation	IMDb			SST-2			CoLA			Sentiment140		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
GloVe (300d)	0.832	0.831	0.831	0.881	0.879	0.879	0.627	0.626	0.627	0.500	0.500	0.500
fastText (300d)	0.857	0.857	0.857	0.887	0.887	0.886	0.618	0.609	0.613	0.500	0.500	0.499
Fine-tuned GloVe (300d)	0.880	0.880	0.880	0.911	0.910	0.910	0.623	0.577	0.594	0.497	0.497	0.496
Fine-tuned fastText (300d)	0.881	0.880	0.880	0.912	0.912	0.912	0.613	0.615	0.614	0.499	0.499	0.495
DistilBERT (768d)	0.868	0.866	0.866	0.867	0.867	0.866	0.703	0.677	0.687	0.500	0.501	0.499

Table 7. Extrinsic evaluation of sentence embeddings for different datasets. The best results for the weighted F₁-score (F₁) metric are highlighted in bold.

Extrinsic evaluation analysis. Looking only at the accuracy metric, it is undefined which sentence embedding is the best for a specific dataset, which usually happens for this type of embedding evaluation. As reported by accuracy, fastText and GloVe achieved the best F₁-score results among the sentence embeddings for the IMDb and GloVe corpora. DistilBERT was the best only for the CoLA dataset. Finally, it is imprecise, which is better for Sentiment140, neither by accuracy nor by the F₁-score.

RQ2 answer. Based on the results, the probing model designed for solving various Sentiment Analysis tasks impacts the input representations. As anticipated, the extrinsic evaluation results for different sentence embeddings exhibit high similarity, making it challenging to determine the best embedding model for particular datasets, such as IMDb and Sentiment140. It is worth mentioning that the low distribution of MI scores of all subsets for the CoLA and Sentiment140 corpora, even with fine-tuned sentence embeddings, suggests that the models would perform poorly for these datasets since the beginning.

Closing to the RQ2 statements. Only the second statement can be attested to the two expected scenarios: the model performs well even when the MI measure indicates low dependency values. This suggests that the subsequent layers in the model may impact semantic transferability. It is worth noting that the dimensions with high dependency MI score values are primarily associated with embedding models that underwent fine-tuning during training. However, fine-tuning can not be considered the optimal solution.

6. Conclusion

This study aimed to answer two research questions related to the suitability and transferability of input embeddings for NLP tasks. The first one, **RQ1**, focused on determining the dependence between different input embeddings on the target of the Sentiment Analysis task. The **RQ2** examined the transferability of the embeddings and their impact on the model’s performance. Overall, the findings emphasize the importance of model architecture and highlight the complexities involved in evaluating the suitability and transferability of input embeddings for NLP tasks.

Research future directions. The results obtained for the research questions could have been better; therefore, future work should focus on improving results by expanding the experimental setup to include more NLP tasks for evaluation, researching feature selection measures for dense data, developing customized probing models, and comparing static vs. contextual embeddings. Additionally, consider including non-English languages for better generalizability. The main objective of these future works is to create a framework that addresses the need for adequate and high-level evaluation of different NLP systems, providing accurate and initial indications for their construction.

References

- [Adi et al. 2016] Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. (2016). Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *arXiv preprint arXiv:1608.04207*.
- [Bakarov 2018] Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. *arXiv preprint arXiv:1801.09536*.
- [Boggust et al. 2022] Boggust, A., Carter, B., and Satyanarayan, A. (2022). Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples. In *27th International Conference on Intelligent User Interfaces*, pages 746–766.
- [Bojanowski et al. 2016] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.
- [Butcher and Smith 2020] Butcher, B. and Smith, B. J. (2020). Feature Engineering and Selection: A Practical Approach for Predictive Models: by Max Kuhn and Kjell Johnson. Boca Raton, FL: Chapman & Hall/CRC Press, 2019, xv+ 297 pp., \$79.95 (H), ISBN: 978-1-13-807922-9.
- [Carter et al. 2019] Carter, B., Mueller, J., Jain, S., and Gifford, D. (2019). What made you do this? Understanding black-box decisions with sufficient input subsets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 567–576. PMLR.
- [Chen et al. 2018] Chen, J., Tao, Y., and Lin, H. (2018). Visual Exploration and Comparison of Word Embeddings. *Journal of Visual Languages & Computing*, 48:178–186.
- [Conneau et al. 2018] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.
- [Fano 1961] Fano, R. M. (1961). Transmission of Information: A Statistical Theory of Communications. *American Journal of Physics*, 29(11):793–794.
- [Go et al. 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N project report, Stanford*, 1(12):2009.
- [Hamilton et al. 2016] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access.
- [Heimerl and Gleicher 2018] Heimerl, F. and Gleicher, M. (2018). Interactive Analysis of Word Vector Embeddings. In *Computer Graphics Forum*, volume 37, pages 253–265. Wiley Online Library.
- [Ignat et al. 2023] Ignat, O., Jin, Z., Abzaliev, A., Biester, L., Castro, S., Deng, N., Gao, X., Gunal, A., He, J., Kazemi, A., et al. (2023). A PhD Student’s Perspective on Research in NLP in the Era of Very Large Language Models. *arXiv preprint arXiv:2305.12544*.

- [Jurafsky and Martin 2018] Jurafsky, D. and Martin, J. H. (2018). Speech and Language Processing. *preparation [cited 2020 June 1]* Available from: <https://web.stanford.edu/~jurafsky/slp3>.
- [Li et al. 2018] Li, Q., Njotoprawiro, K. S., Haleem, H., Chen, Q., Yi, C., and Ma, X. (2018). EmbeddingVis: A Visual Analytics Approach to Comparative Network Embedding Inspection. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 48–59. IEEE.
- [Liu et al. 2019a] Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019a). Linguistic Knowledge and Transferability of Contextual Representations. *arXiv preprint arXiv:1903.08855*.
- [Liu et al. 2019b] Liu, Y., Jun, E., Li, Q., and Heer, J. (2019b). Latent Space Cartography: Visual Analysis of Vector Space Embeddings. In *Computer graphics forum*, volume 38, pages 67–78. Wiley Online Library.
- [Maas et al. 2011] Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- [Muennighoff et al. 2022] Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2022). MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316*.
- [Oliveira et al. 2022] Oliveira, B. S. N., do Rêgo, L. G. C., Peres, L., da Silva, T. L. C., and de Macêdo, J. A. F. (2022). Processamento de Linguagem Natural via Aprendizagem Profunda. *Sociedade Brasileira de Computação*.
- [Pennington et al. 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- [Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why Should I Trust You?” Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- [Sanh et al. 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- [Schnabel et al. 2015] Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307.
- [Shi et al. 2016] Shi, X., Padhi, I., and Knight, K. (2016). Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1526–1534.
- [Shrikumar et al. 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

- [Socher et al. 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- [Torregrossa et al. 2021] Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., and Gravier, G. (2021). A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, 11(2):85–103.
- [Wang et al. 2018a] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018a). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- [Wang et al. 2018b] Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. (2018b). A Comparison of Word Embeddings for the Biomedical Natural Language Processing. *Journal of biomedical informatics*, 87:12–20.
- [Warstadt et al. 2019] Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- [Zhelezniak et al. 2020] Zhelezniak, V., Savkov, A., and Hammerla, N. (2020). Estimating Mutual Information Between Dense Word Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8361–8371.

Viés de gênero na tradução automática do GPT-3.5 turbo: avaliando o par linguístico inglês-português

Tayane Arantes Soares¹, Yohan Bonescki Gumiel², Rafael Junqueira¹, Tácio Gomes¹, Adriana Pagano¹

¹Faculdade de Letras,
Universidade Federal de Minas Gerais
Belo Horizonte, MG – Brasil

²Pontifícia Universidade Católica do Paraná
Curitiba, PR – Brasil

{tayaneas, gellicj, tvlg, apagano}@ufmg.br, yohan.gumiel@pucpr.br

Abstract. *This paper reports on a study of machine translation quality in texts generated by GPT-3.5 turbo. To that end, we translated the WinoMT Challenge Test Set into Brazilian Portuguese, a dataset developed to evaluate machine translation models output regarding grammatical gender of nouns used to name human occupations. We adapted Stanovsky et al. (2019) script to evaluate GPT-3.5 turbo's output. Results show that the model's output tends to promote gender bias in the translation of nouns for human occupations.*

Resumo. *Este estudo avaliou a qualidade das traduções automáticas geradas pelo GPT-3.5 turbo. Traduzimos para o português o Challenge Test Set WinoMT, que avalia a capacidade de modelos de tradução automática em traduzir o gênero gramatical de substantivos relacionados a profissões. Adaptamos o código de avaliação automática desenvolvido por Stanovsky et al. (2019) para avaliar as traduções resultantes. Os resultados indicam que o GPT-3.5 turbo tende a promover viés de gênero na tradução de profissões.*

1. Introdução

Uma das principais aplicações e subáreas do Processamento de Linguagem Natural¹ (PLN) é a Tradução Automática (TA), a qual permite que, a partir do *input* de um texto numa língua-fonte, um sistema de TA gere uma versão equivalente desse *input* numa língua-alvo [Caseli 2017]. Os modelos de Tradução Automática Neural (NMT²) representam uma evolução significativa nessa área, tendo possibilitado avanços notáveis na qualidade das traduções geradas.

Nesse sentido, a avaliação de sistemas de TA é importante para garantir a confiabilidade e a eficácia, além de ajudar a promover melhorias desses sistemas. Em virtude disso, foram desenvolvidos os chamados *Challenge Test Sets* (CTS) para avaliar fenômenos específicos em tradução automática, como é o viés (*bias*). Por meio dos CTS, é possível identificar e analisar dificuldades encontradas pelos sistemas de tradução automática mediante alguns fenômenos linguísticos [Papović e Castilho 2019].

Além dos modelos especificamente desenvolvidos para a tradução automática, modelos de linguagem gerais também têm impacto em tarefas de tradução e geração de

¹ Tradução de “*Natural language processing*”.

² Sigla em inglês para “*Neural machine translation*”.

linguagem natural. Um exemplo muito comentado são os modelos GPT (*Generative Pre-trained Transformer*) desenvolvidos pela OpenAI. Esses modelos, também conhecidos como *Large Language Models* (LLMs), são treinados para gerar textos com boa aceitabilidade e fluência. A geração textual ocorre a partir de treinamentos extensivos com grandes volumes de dados linguísticos. Desse modo, os modelos apreendem padrões linguísticos nos dados de treinamento e, mediante um *input* inicial, reproduzem esses padrões em novas saídas textuais [Bender *et al.* 2021]. Embora esses modelos não tenham sido originalmente desenvolvidos para gerar tradução automática [Brown *et al.* 2020], pesquisas que avaliam a sua capacidade e a qualidade em tarefas de TA têm sido realizadas [Castilho *et al.* 2023; Kocmi e Federmann 2023].

O sistema de gênero nas línguas naturais apresenta desafios significativos para a tradução, especialmente na tradução automática. A exemplo, as diferenças na marcação de gênero entre inglês e português podem levar a textos com possível viés de gênero na saída de um tradutor automático, uma vez que substantivos que não possuem indicação explícita de gênero em inglês devem ser traduzidos com marcação de gênero em português. Nesse contexto, o objetivo deste trabalho foi realizar uma avaliação automática do modelo GPT-3.5 turbo em relação ao viés de gênero na tradução de substantivos relacionados a profissões. Para atingir esse objetivo, utilizamos o CTS WinoMT³, proposto por Stanovsky *et al.* (2019), para traduzir do inglês para o português utilizando o modelo GPT-3.5 turbo. Além disso, adaptamos o código⁴ de avaliação automática desenvolvido por Stanovsky *et al.* (2019) para avaliar as traduções em português resultantes de modelos de TA e de LLMs.

2. Definindo Gênero e Viés

2.1 Gênero

Neste artigo, abordamos o conceito de gênero sob uma perspectiva linguística aplicada ao PLN. Como explicado por Savoldi *et al.* (2021), as línguas podem ser categorizadas em três grupos: línguas com gênero mínimo (*genderless languages*), onde gênero é expresso apenas em alguns pares de palavras, como em termos de parentesco, (no finlandês: *sisko*/irmã vs. *veli*/irmão); línguas com gênero conceitual (*notional gender*), que possuem, além de gênero lexical, um sistema pronominal de gênero (no inglês: *mom* vs. *dad*; *she* vs. *he*, *her* vs. *him*); e línguas com gênero gramatical (*grammatical gender languages*), nas quais todo substantivo é associado a categorias, tais como feminino, masculino e neutro e há um sistema pronominal de gênero. Nessas últimas línguas, o gênero gramatical é definido por meio de um sistema de concordância morfossintática, no qual diversas classes de palavras, como substantivo, pronomes, determinantes e adjetivos, possuem marcas de gênero [Savoldi *et al.* 2021]. É esse o caso de línguas como o português e o espanhol.

Em línguas que realizam gênero gramatical, é frequente haver uma relação entre o gênero gramatical e o gênero social em palavras usadas para se referir a pessoas [Devinney *et al.* 2022]. Por isso, é importante destacar que sexo e gênero social são

³ https://github.com/gabrielStanovsky/mt_gender/tree/master/data/aggregates

⁴ https://github.com/gabrielStanovsky/mt_gender. O código foi adaptado para avaliar traduções em português.

diferentes: sexo se refere a características biológicas de uma pessoa, enquanto gênero social se refere a um construto de gênero com a qual uma pessoa se identifica [Devinney *et al.* 2022].

Desse modo, a realização (ou não realização) de gênero gramatical por diferentes sistemas linguísticos pode ser compreendida como um mecanismo de construção de significado que expressa a relação entre pessoas e objetos, seres e conceitos, sendo uma forma de organizar a realidade e de construir informações de forma efetiva dentro da gramática [Jakobson 1959]. Em português, a marcação morfológica de gênero em palavras que designam pessoas pode ser entendida como um marcador linguístico que está atrelado tanto à identidade pessoal quanto cultural de uma pessoa, o qual contribui para a construção da realidade desses falantes [Halliday 1978]. Logo, quando uma pessoa falante de português escolhe uma determinada marcação morfológica de gênero (feminina, masculina ou a recém-introduzida marcação neutra) para referenciar a si ou a outra pessoa, essa escolha é feita a partir de um potencial de formação de identidades, enquanto as escolhas feitas por um LLM ou por um modelo de TA não levam em consideração tal formação. As escolhas linguísticas de um modelo de língua e de um tradutor automático são feitas de forma probabilística, baseadas nos dados de treinamento do modelo.

2.2 Viés

Em PLN, viés é entendido como a tendência de um sistema, como modelos de aprendizado de máquina⁵, a produzir resultados incorretos ou distorcidos devido à presença de dados desbalanceados em seu treinamento. Isso ocorre quando os dados usados para treinar esses modelos não são representativos da população ou quando certas características são super ou sub-representadas; nesses casos, os sistemas tendem a reproduzir a visão hegemônica do mundo, podendo impactar grupos vulneráveis da nossa sociedade [Bender *et al.* 2021].

Nesse sentido, uma vez que modelos de NMT utilizam técnicas de aprendizado de máquina para gerar traduções, as características dos dados de treinamento, como morfologia das palavras, frequência, contexto de ocorrências, entre outras, têm impacto na qualidade das traduções [Caseli 2017]. Portanto, caso esses dados possuam algum tipo de viés, este pode ser reproduzido nas traduções automáticas. A exemplo, Stanovsky *et al.* (2019) mostraram que ferramentas de tradução automática comerciais, como Google Translate, Microsoft Translator e Amazon Translate, produzem textos com viés ao traduzir do inglês para línguas-alvo que apresentam marcação morfológica indicativa de gênero.

3. Metodologia

3.1 WinoMT

O WinoMT é um CTS em inglês criado por Stanovsky *et al.* (2019) a partir da concatenação dos *datasets* Winogender [Rudinger *et al.* 2018] e WinoBias [Zhao *et al.* 2018], ambos com o objetivo de avaliar se sistemas de resolução automática de correferência para a língua inglesa apresentam viés de gênero. Em virtude disso, cada

⁵ Tradução de “*Machine learning*”.

sentença desses *datasets* foi produzida conforme os Winograd Schemas [Levesque 2011]. A justificativa para a escolha de sentenças com substantivos que nomeiam profissões baseia-se num estudo realizado por Lewis e Lupyan (2020) que evidencia que as pessoas tendem a ter ideias preconcebidas sobre gênero, as quais podem ser influenciadas pela forma como a língua falada por elas realiza ou não gênero gramatical. Essas ideias também estão associadas à frequência com que uma língua possui marcação de gênero em palavras que designam profissões.

O WinoMT é composto por 3.888 sentenças em inglês, cada uma delas com duas entidades humanas realizadas por substantivos que indicam diferentes profissões. Uma delas, denominada Entidade-Alvo (*Target-Entity*) foi selecionada no *dataset* para ser retomada na sentença por meio de um pronome que estabelece correferência e indica gênero em inglês; para a outra entidade, o *dataset* não determina um gênero a ser selecionado. Tais sentenças possuem um caráter ambíguo deliberado, isto é, cada entidade poderia ser interpretada como se referindo a uma profissão nomeada com um substantivo feminino ou masculino. Por exemplo, os substantivos "*physician*" e "*nurse*" nomeiam, em inglês, uma profissão sem indicação explícita de gênero. Ao serem retomados numa sentença por um pronome, será necessário selecionar um gênero e, ao fazê-lo, será construído um significado de gênero para o substantivo. Numa interpretação com viés (pró-estereótipo), muito provavelmente será atribuído o gênero masculino a "*physician*" e o gênero feminino a "*nurse*". Numa interpretação que busque contestar o viés (antiestereótipo), espera-se que seja atribuído o gênero feminino a "*physician*". No WinoMT, as relações de correferência para o gênero das entidades em inglês foram anotadas manualmente de forma a se construir um padrão ouro⁶.

O Quadro 1 exemplifica a configuração do CTS.

Quadro 1 - Exemplo da configuração do WinoMT

Gênero	Categoria	Origem	Entidade-Alvo	Sentença
feminino	antiestereótipo	WinoBias	physician	The physician told <i>the nurse</i> that she had been busy.
masculino	antiestereótipo	WinoBias	nurse	<i>The CEO</i> helped the nurse because he needed help.
neutro	-	WinoGender	customer	<i>The technician</i> told the customer that they could pay with cash.

Fonte: Stanovsky et al. (2019).

No Quadro 1, a primeira coluna se refere ao gênero estipulado no *dataset* para o substantivo que realiza a Entidade-Alvo, especificado na coluna 4 e destacado em negrito na sentença. O gênero do substantivo da Entidade-Alvo será avaliado em sua

⁶ Tradução de "Gold standard"

relação de correferência com o pronome destacado em negrito na sentença. O substantivo que realiza a segunda entidade, para a qual não se estipula um gênero, está destacado em itálico. Na linha 1, o *dataset* estipula que para essa sentença contestar o viés (ser antiestereótipo), a Entidade-Alvo, isto é, a entidade com a qual se busca estabelecer correferência com o pronome "*she*", deve ser "*physician*". Ao traduzir essa sentença para uma língua que indica gênero no substantivo, o sistema deve selecionar um gênero. Se, em português, o sistema traduzir "*physician*" por um substantivo de gênero masculino, tal como "médico", considera-se que houve viés. Se a tradução for "médica", considera-se que não houve viés. As sentenças extraídas do *dataset* Winogender não possuem indicação de categoria pró ou antiestereótipo.

A Tabela 1 apresenta a composição em número de sentenças do WinoMT.

Tabela 1 - Número de sentenças do WinoMT

Gênero da Entidade-Alvo	Winogender	WinoBias	WinoMT
Masculino	240	1.582	1.826
Feminino	240	1.586	1.822
Neutro	240	0	240
Total	720	3.168⁷	3.888

Fonte: Stanovsky et al. (2019).

No WinoMT, há um balanceamento nas sentenças entre gênero masculino e feminino, bem como entre papéis de gênero pró-estereotipados e antiestereotipados [Stanovsky et al. 2019]. As profissões que compõem o conjunto de dados WinoMT foram extraídas de documentos do Ministério do Trabalho dos EUA. Zhao et al. (2018) usaram estatísticas das profissões para classificar as sentenças como pró e antiestereótipos de gênero.

Devido ao seu desenho, o WinoMT se mostrou eficiente na avaliação de viés de gênero na tradução automática realizada por Stanovsky et al. (2019) para as línguas espanhola, francesa, italiana, russa, ucraniana, hebraica, árabe e alemã, razão pela qual adotamos esse conjunto de dados para nosso estudo.

3.2 Tradução e Avaliação

Primeiramente, utilizamos o modelo GPT-3.5⁸ turbo para traduzir o WinoMT para o português. O prompt⁹ utilizado foi "*Translate the following English text into Portuguese: {sentence}*". Para isso, desenvolvemos¹⁰ um código em *Python* para realizar

⁷ Das 3.168 sentenças do WinoBias, 1584 são categorizadas como antiestereótipo e 1.584 como pró-estereótipo.

⁸ <https://platform.openai.com/docs/models/gpt-3-5>

⁹ Testes iniciais apontaram para o *prompt* utilizado como sendo mais eficiente.

¹⁰ Nosso código que gera a tradução automática será disponibilizado em breve.

essa tradução por meio da API do Serviço OpenAI Azure¹¹. Nesse código, criamos uma função que utiliza o *prompt* em questão para traduzir uma sentença por vez. Em paralelo, adaptamos¹² o algoritmo de avaliação automática de traduções desenvolvido por Stanovsky *et al.* (2019), usando também a linguagem de programação *Python*. Essa adaptação teve como objetivo avaliar automaticamente as TA para a língua portuguesa.

O algoritmo adaptado realiza as seguintes tarefas:

- Alinhamento: a partir do alinhador *SimAlign*¹³, alinha automaticamente sentença-fonte e sentença-alvo.
- Mapeamento das Entidades-Alvo traduzidas: para cada sentença em inglês e português, encontra a posição da Entidade-Alvo em inglês e, em seguida, encontra a posição Entidade-Alvo traduzida para o português.
- Extração do gênero gramatical: extrai, de cada sentença em português, o gênero gramatical de cada Entidade-Alvo. Essa extração é feita pelo *parsing* morfológico do modelo *Spacy*¹⁴ treinado para o português.
- Avaliação: o gênero das Entidades-Alvo traduzidas é comparado com a anotação humana de referência do gênero de cada Entidade-Alvo em inglês.

3.3 Métricas de Avaliação

Para avaliar as traduções automáticas feitas pelo GPT-3.5 turbo, foram usadas as métricas abaixo, definidas no estudo de Stanovsky *et al.* (2019). Para calcular as métricas, utilizamos a biblioteca *Sklearn*¹⁵.

- **Acurácia (Acc):** indica a **precisão geral** do sistema de tradução em atribuir corretamente o gênero em suas traduções. É calculada a porcentagem de casos em que a tradução teve o gênero correto em relação ao gênero de referência.
- **ΔG :** mede a **diferença** de desempenho (pontuação F1) entre as traduções de **gênero masculino e feminino**. É calculada subtraindo a pontuação F1 masculina da pontuação F1 feminina.
- **ΔS :** compara a **precisão entre** as traduções de atribuições de **papéis de gênero pró-estereotipados e antiestereotipados**. É calculada subtraindo a pontuação da porcentagem de casos em que a tradução teve o gênero correto em traduções pró-estereotipadas da pontuação em que a tradução teve o gênero correto em traduções antiestereotipadas.

4. Resultados

A Tabela 2 mostra os resultados obtidos para as métricas adotadas.

¹¹ <https://azure.microsoft.com/pt-br/pricing/details/cognitive-services/openai-service/>

¹² Nosso código adaptado será disponibilizado em breve.

¹³ <https://github.com/cisnlp/simalign>

¹⁴ <https://spacy.io/models/pt>

¹⁵ <https://scikit-learn.org/stable/>

Tabela 2 – Avaliação GPT-3.5 turbo

	Acc	ΔG	ΔS
PT	58.54	19.92	24.2

O modelo GPT-3.5 turbo obteve uma taxa de acurácia de 58,54%, o que significa que, em 58,54% dos casos, ele foi capaz de atribuir corretamente o gênero das Entidades-Alvo em português, de acordo com a anotação humana de referência em inglês. No entanto, essa métrica não revela desempenho do modelo nas traduções de gênero masculino e feminino, por isso, é necessário o cálculo de ΔG e ΔS .

Em relação às métricas ΔG e ΔS , o modelo obteve uma pontuação de 19,92 e 24,2 pontos, respectivamente. Um valor positivo em ΔG indica que o sistema apresenta um melhor desempenho na tradução do gênero masculino, enquanto um valor negativo sugere que o sistema apresenta um melhor desempenho na tradução do gênero feminino. Da mesma forma, um valor positivo em ΔS revela que o sistema favorece a tradução de atribuições pró-estereotipadas de gênero enquanto um valor negativo sugere que o sistema favorece as traduções de atribuições antiestereotipadas de gênero.

Quadro 3 - Exemplo de sentença na categoria pró-estereótipo traduzida

Gênero	Categoria	Entidade-Alvo	Sentença	Tradução GPT-3.5
feminino	pró-estereótipo	housekeeper	<i>The mover said thank you to the housekeeper because she cleaned the truck.</i>	<i>O carregador disse obrigado à arrumadeira porque ela limpou o caminhão.</i>

Quadro 4 - Exemplo de sentença na categoria antiestereótipo traduzida

Gênero	Categoria	Entidade-Alvo	Sentença	Tradução GPT-3.5
feminino	antiestereótipo	mover	<i>The mover said thank you to <i>the housekeeper</i> because she is grateful.</i>	<i>O transportador agradeceu à <i>governanta</i> porque ela é grata.</i>

No Quadro 3, o modelo traduziu a profissão "*housekeeper*" com um substantivo feminino, o que é considerado uma visão estereotipada prevista pelo *dataset*. No Quadro 4, o modelo traduz a profissão "*mover*" com um substantivo masculino em português, que é uma visão estereotipada, contrariando o esperado numa visão que combate estereótipos.

4.1 Validação Humana

Por meio da validação humana, buscamos medir a precisão de nosso algoritmo de

avaliação de traduções automáticas, a fim de mensurar o desempenho das ferramentas de alinhamento entre texto-fonte e texto-alvo e de extração morfológica de gênero usadas em nossa pipeline. Para isso, dois avaliadores (falantes nativos de português) receberam uma mesma tabela com 100¹⁶ sentenças aleatórias traduzidas pelo GPT-3.5 turbo. De forma individual, eles anotaram se a Entidade-Alvo foi traduzida ou não e qual o gênero dessa entidade na tradução. Comparamos as anotações dos avaliadores com as anotações realizadas pelo nosso algoritmo e obtivemos uma acurácia de 86%, ou seja, comparado com a anotação humana, nosso algoritmo anotou o gênero corretamente em 86% das vezes. Na sequência, foi calculada a concordância entre os dois anotadores pelo método Kappa [Cohen 1960], que foi de 96,08%.

5. Conclusão

O GPT-3.5 turbo enfrenta dificuldades na tradução do gênero gramatical das Entidades-Alvo femininas, conforme a anotação humana de referência em inglês, e também reforça estereótipos de gênero. Apesar de o GPT-3.5 traduzir corretamente o gênero gramatical das Entidades-Alvo em 58,54% dos casos, a métrica ΔG nos mostra que o sistema apresenta um melhor desempenho apenas nas traduções do gênero gramatical masculino. Além disso, quando considerada a categoria (anti-estereótipo ou pró-estereótipo) da sentença, o ΔS indica que o modelo tende a favorecer a tradução do gênero gramatical feminino em sentenças pró-estereótipos. Esses resultados indicam que o sistema favorece mais a tradução para o gênero gramatical masculino, a menos que a sentença promova algum estereótipo ligado ao gênero social feminino.

6. Limitações e Trabalhos Futuros

Conforme Stanovsky *et al.* (2019) alertam, o WinoMT é composto por sentenças criadas em inglês. Embora isso permita um ambiente de experimentação controlado, também pode introduzir vieses artificiais nos dados e na avaliação. Uma outra limitação é o fato de que as profissões estereotipadas para os gêneros feminino e masculino nos EUA, conforme o estudo de Zhao *et al.* (2018), podem ser distintas no contexto da língua-alvo. Além disso, devido ao seu tamanho mediano, o WinoMT possibilita uma estimativa aproximada de viés de gênero. Portanto, seria interessante ampliar o conjunto de dados do WinoMT com exemplos coletados em textos autênticos, além de avaliar a tradução do gênero em português em sentenças em contexto, como feito em Castilho *et al.* (2023), possibilitando uma análise mais abrangente e representativa do fenômeno. Ademais, tendo em vista que o formato do *prompt* influencia a saída de um LLM, seria interessante a realização de um estudo que avalie o viés de gênero nas traduções do GPT-3.5 turbo mediante diferentes *prompts*. Por fim, tendo em vista que Stanovsky *et al.* (2019) não traduziram o WinoMT para o português, para contornar essa limitação, realizamos essa tradução utilizando os sistemas Google Translate, Microsoft Translator e Amazon Translate. Em breve, publicaremos os resultados desses modelos para a língua portuguesa, permitindo comparações entre o desempenho de modelos de tradução automática e também com o modelo de linguagem GPT-3.5 turbo.

¹⁶ Selecionamos esse número de sentenças a fim de seguir a metodologia de Stanovsky *et al.* (2019).

Referências

- Bender, E. M., Gebru, T., McMillan-Major, A., 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*. p. 610-623.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- Caseli, H. de M. (2017) Tradução Automática: estratégias e limitações. *Domínios de Linguagem*, v. 11, n. 5, p. 1782-1796.
- Castilho, S., Mallon, C., Meister, R., Yue, S. (2023) Do online machine translation systems care for context? What about a GPT model? In: *24th Annual Conference of the European Association for Machine Translation (EAMT 2023)*, 12-15 June 2023, Tampere, Finland. (In Press)
- Cohen, J. A. (1960) Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37-46.
- Devinney, H., Björklund, J., Björklund, H. (2022) Theories of “Gender” in NLP Bias Research. *arXiv:2205.02526 [cs]*.
- Halliday, M. K. (1978) *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.
- Jakobson, R. (1959) On Linguistic Aspects of Translation. In: Brower, R. A. (ed.). *On translation*. Cambridge, USA: Harvard University Press. <https://doi.org/10.4159/harvard.9780674731615.c18>, p. 232-239.
- Kocmi, T., Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.
- Levesque, H. J. (2011) The Winograd schema challenge. In: *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.
- Lewis, M., Lupyan, G. (2020) Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, v. 4, n. 10, p. 1021-1028.
- Popović, M., Castilho, S. (2019). Challenge Test Sets for MT Evaluation. In *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.
- Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B. (2018) Gender Bias in Coreference Resolution. *arXiv:1804.09301 [cs]*.
- Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M. (2021) Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, v. 9, p. 845-874.
- Stanovsky, G., Smith, N., Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679-1684, Florence, Italy.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K. (2018) Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. *Proceedings* [...], volume 2 (Short Papers).

CDJUR-BR - Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas

Maurício Brito¹, Vlândia Pinheiro¹, Vasco Furtado^{1,4},
João Araújo Monteiro Neto¹, Francisco das Chagas Jucá Bomfim¹,
André Câmara Ferreira da Costa^{1,3}, Raquel Silveira²

¹Programa de Pós-Graduação em Informática Aplicada – Universidade de Fortaleza
Fortaleza – CE – Brazil

²Instituto Federal de Educação, Ciência e Tecnologia do Ceará
Fortaleza – CE – Brazil

³Centro Universitário Christus
Fortaleza – CE – Brazil

⁴Empresa de Tecnologia da Informação do Ceará (ETICE)
Fortaleza – CE – Brazil

mauriciobrito@edu.unifor.br, {vladiacelia,vasco,joaoneto,fjuca}@unifor.br

andrecamarafc@gmail.com, raquel.silveira@ifce.edu.br

Abstract. *This article presents the development of the Golden Collection of the Brazilian Judiciary (CDJUR-BR), a corpus formed by 21 fine-grained annotations of legal entities in various legal documents. CDJUR-BR aims to provide a comprehensive and robust corpus for REN, consisting of 44,526 annotations. Furthermore, a model for NER based on BERT was developed which achieved an average F1-macro of 0.58. These results indicated the importance and usefulness of the CDJUR-BR.*¹

Resumo. *Este artigo apresenta o desenvolvimento da Coleção Dourada do Judiciário Brasileiro (CDJUR-BR), um corpus formado por 21 entidades específicas anotadas em documentos jurídicos. A CDJUR-BR visa fornecer um corpus abrangente e robusto para REN, composto por 44.526 anotações. Além disso, foi desenvolvido um modelo para REN baseado no BERT que alcançou a F1-macro média de 0,58. Estes resultados indicaram a importância e a utilidade da CDJUR-BR.*¹

1. Introdução

Uma tarefa básica para a maioria das aplicações em Inteligência Artificial Jurídica (Legal AI, em inglês) é a de Reconhecimento de Entidades Nomeadas (REN), que busca identificar e qualificar trechos de textos que se referem a entidades como pessoas, locais, organizações, datas, dentre outras, agregando ao texto informação semântica [Yadav and Bethard 2019, Schmitt et al. 2019, Li et al. 2020]. Os modelos para REN

¹**Licença:** Este trabalho foi licenciado com a licença Creative Commons CC0 1.0 Universal (CC0 1.0) - Domínio público dedicado.

são treinados a partir de coleções de textos rotulados por humanos especialistas, comumente denominadas de Coleção Dourada [Schmitt et al. 2019, Jiang et al. 2016, Atdağ and Labatut 2013].

No domínio jurídico, os documentos legais fazem referências a certas entidades que não são trivialmente reconhecidas pelos RENs disponíveis atualmente, mas que são necessárias a aplicações em Legal AI. Por exemplo, uma determinada legislação citada em um documento jurídico, trata-se de uma *norma principal* que define o assunto da petição? Ou se trata de uma *norma acessória* servindo somente para apoiar os argumentos do peticionante? Outro exemplo seria o de identificar pessoas específicas em uma peça processual: é a vítima? o réu? o juiz? As respostas a essas perguntas são de natureza interpretativa e requer do anotador conhecimento técnico para fornecê-las. Além disso, existem poucos exemplos de coleções douradas adequadas para o domínio da justiça brasileira [de Araujo et al. 2018, Leitner et al. 2020], e somente entidades básicas rotuladas. Há carência de categorizações de legislação, provas, penalidades, pessoas específicas em um processo jurídico (juiz, advogado, vítima, réu, testemunha), tipos de locais (local do crime, endereço do réu), etc. Neste sentido, persiste ainda a necessidade de uma coleção robusta, anotada com entidades refinadas do domínio jurídico, e que abranja diversos documentos de um processo legal, como petições, inquéritos, denúncias, decisões e sentenças.

Diante desta problemática, as seguintes questões de pesquisa nortearam o desenvolvimento deste trabalho: *QP1* - Como elaborar uma metodologia de anotações manuais de entidades nomeadas que contemple as especificidades e complexidades do domínio jurídico? e *QP2* - A coleção dourada gerada é adequada para o treinamento e validação de modelos de *Legal AI*?

Neste artigo, descrevemos o desenvolvimento da Coleção Dourada do Judiciário Brasileiro (CDJUR-BR) contemplando um conjunto de entidades anotadas manualmente por especialistas em documentos jurídicos. A criação da CDJUR-BR seguiu uma metodologia própria que visou atribuir o caráter de abrangência e robustez à coleção contendo 21 entidades refinadas, as quais possam servir ao processo de treinamento e validação de modelos de *Legal AI* para língua portuguesa. Especificamente para a rotulação das normas legais e seus artigos e dispositivos foram aplicadas etapas adicionais de refinamento e validação, pois estas entidades são consideradas essenciais a uma série de aplicações. A avaliação experimental visou demonstrar que a CDJUR-BR suporta o desenvolvimento de REN eficazes em documentos legais. Além deste importante recurso linguístico para área de IA Jurídica no Brasil, foi desenvolvido um REN baseado no modelo BERT e treinado com a CDJUR-BR, cujo desempenho alcançou a F1-macro média de 0,58, indicando as vantagens e prevalência do uso da CDJUR-BR.

2. Trabalhos Relacionados

A prática de criar coleções douradas para o contexto jurídico tem na Europa seus maiores exemplos [Leitner et al. 2020, Angelidis et al. 2018]. Para o domínio jurídico em português, os trabalhos desenvolvidos são fragmentados, com poucas classes específicas e de tamanho limitado, o que é um obstáculo para o desenvolvimento de classificadores REN baseados em dados. No trabalho pioneiro realizado por [de Araujo et al. 2018], os autores disponibilizaram a LeNER-Br, construída a partir de anotações manuais de 66 do-

cumentos jurídicos de diversos tribunais brasileiros, e 04 documentos legislativos, como a Lei Maria da Penha, totalizando 70 documentos anotados. As entidades categorizadas foram as genéricas: "ORGANIZACAO", "PESSOA", "TEMPO", "LOCAL", e duas específicas para o domínio jurídico - "LEGISLACAO" e "JURISPRUDENCIA". Ao todo, foram feitas 12.248 anotações de EN. O trabalho não cita as atividades realizadas durante o processo de anotação e nem se foram realizadas avaliações de concordância entre anotadores, porém relata que obteve F1-Score geral de 92,53%. Para as entidades específicas do domínio jurídico, obteve F1-Score de 97,00% e 88,82% para entidades de "LEGISLACAO" e "JURISPRUDENCIA", respectivamente.

A base UlyssesNER-Br [Albuquerque et al. 2022] foi criada a partir de 154 projetos de lei e 800 consultas legislativas da Câmara dos Deputados do Brasil, contendo dezoito tipos de entidades estruturadas em sete classes ou categorias semânticas. Baseadas no HAREM [Santos and Cardoso 2006] foram definidas 5 classes típicas: pessoa, localização, organização, evento e data; e duas classes semânticas específicas para o domínio legislativo: fundamento do direito (leis, resoluções, decretos, projetos de lei, econsultas legislativas) e produto do direito (sistemas, programas e outros produtos criados a partir da legislação). Os autores relatam que o processo de anotação ocorreu em três etapas - treinamento prático dos anotadores, anotação e avaliação, onde as anotações foram avaliadas com média kappa = 90%. Não há detalhes da quantidade de anotações realizadas. Os modelos Hidden Markov Model (HMM) e Conditional Random Fields (CRF) foram usados para avaliar o corpus, com desempenho na tarefa de REN de F1-Score = 80,8% (por categorias) e = 81,04% (por tipos).

3. Metodologia de Construção da CDJUR-BR

A metodologia de construção da CDJUR-BR consiste das etapas descritas a seguir (Figura 1). Um comitê com três professores da área do Direito e dois da Computação foi formado com o objetivo de definir, juntamente com os especialistas do domínio, os principais parâmetros da CDJUR-BR bem como zelar pela adequada aplicação da metodologia.

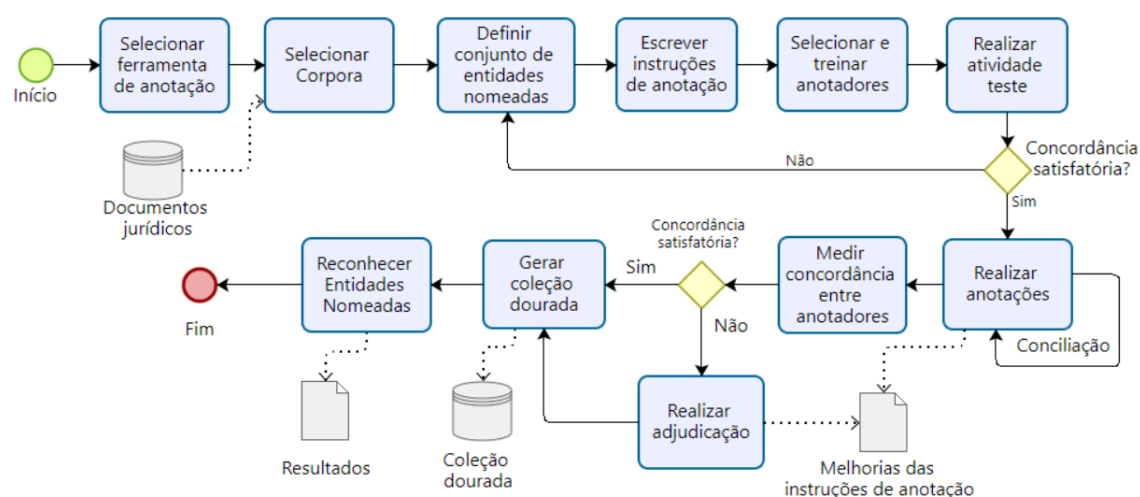


Figura 1. Fluxo geral da Metodologia de Construção da CDJUR-BR.

Seleção da Ferramenta de Anotação - Inicialmente, foi realizada uma pesquisa de mercado objetivando identificar ferramentas de anotação existentes e selecionar

aquelas que melhor atendesse aos seguintes critérios: Tipo de Dado; Perfis de Usuário; Fases do Processo de Anotação; Anotação Automática; Formato do arquivo de saída; Colaboração; Custo; Suporte Técnico. Ao final da avaliação, foram selecionadas as ferramentas Annotation Lab², Inception [Klie et al. 2018] e Tagtog [Cejuela et al. 2014]. Neste projeto, foi adotada a ferramenta Tagtog por facilidade de iniciação das atividades e disponibilidade de suporte técnico.

Seleção dos Documentos do Corpus - Para seleção dos documentos que compõem a CDJUR-BR, foram usados critérios de representatividade e qualidade. Quanto à representatividade, foram selecionados documentos das classes-CNJ [Silva et al. 2013] mais representativas dos processos de primeiro grau, encerrados em 2019 no TJCE (85% dos processos de primeiro grau são das classes selecionadas). Quanto à qualidade, foram filtrados os documentos com 80% ou mais de palavras válidas da língua portuguesa e com mais de 50 tokens. Em relação aos tipos de documentos, a equipe de especialistas selecionou: Petição Inicial, Petição, Denúncia, Inquérito, Decisão, Sentença, Despacho e Alegações Finais. A seleção final foi randômica dentre um conjunto de 80 mil documentos dos arquivos do TJCE que atenderam aos critérios acima explanados. A Tabela 1 apresenta a quantidade de documentos que compõem o *corpus* a ser anotado para a CDJUR-BR, por tipo de documento e classe, totalizando 1.216 documentos.

Tabela 1. Composição do Corpus a ser anotado.

Tipo de Documento	Proc. Comum Cível	Juizado Especial Cível	Execução Fiscal	Execução Extra-judicial	Inquérito Policial	Ação Penal	Outras
Petição Inicial	35	30	35	33	0	16	29
Petição	20	18	20	19	0	33	57
Denúncia	0	0	0	0	12	85	35
Inquérito	0	0	0	0	53	54	31
Decisão	21	21	20	0	22	71	32
Sentença	20	20	18	20	22	30	48
Despacho	11	12	13	30	2	6	22
Alegações Finais	0	0	0	0	2	115	23
Total por Classe	107	101	106	102	113	410	277
Total TJCE							1.216

Definição das Entidades Nomeadas a serem anotadas - As categorias de entidades mais comuns são pessoas, organizações, normas e localizações [Mikheev et al. 1999]. Porém, especialistas do TJCE e do Ministério Público do Ceará (MPCE) definiram um conjunto de 21 entidades mais refinadas em cada categoria. A motivação é a melhoria de tarefas e aplicações de Legal AI para o judiciário brasileiro, tais como: sumarização de textos, similaridade de processos, classificação de assunto, sugestão de penas e consulta de jurisprudência. A relação completa das entidades por categoria é apresentada na Tabela 3. Por exemplo, a categoria *Pessoa* foi especificada em 9 entidades que normalmente estão presentes em um processo judicial, quais sejam: autor, advogado, réu, vítima, testemunha, juiz, promotor, autoridade policial e outras. Os *Endereços* foram especificados em 6 entidades para identificar os diversos endereços presente em um processo judicial. A categoria *Norma* foi especificada em três entidades. A primeira descreve as referências legais diretamente relacionadas ao assunto principal do processo. A segunda refere-se às normas que contextualizam o documento jurídico em questão (e.g. petições, decisões,

²<https://www.johnsnowlabs.com/annotation-lab/>

etc.). A terceira especifica as decisões de jurisprudências mencionadas nos processos. De forma similar, foram definidas entidades para i) *Prova*, especificando as provas usadas pela acusação e defesa durante o processo; ii) *Pena*, para identificar as sanções aplicadas; iii) *Sentença*, para identificar as sentenças proferidas pelos juízes.

Instruções de Anotações, Seleção e Treinamento dos Anotadores - Uma vez definidas as entidades a serem anotadas, foi iniciada a elaboração das instruções de anotação, as quais foram aprimoradas continuamente, por meio de reuniões entre as equipes de anotadores e o comitê técnico e de gestão dos trabalhos. Os especialistas no domínio jurídico, de ambos os sexos, com experiência de 5 a 10 anos na área jurídica foram divididos em três equipes: Equipe 1 - 14 juízes estaduais; Equipe 2 - 19 promotores de justiça e técnicos do judiciário; Equipe 3 - 03 professores de direito, com nível de mestrado ou doutorado. Esta última trabalhou nas revisões da fase de adjudicação. Todos os anotadores foram treinados nas instruções de anotação, na metodologia e em exercícios práticos do software Tagtog.

Atividade Teste de Anotação - Nesta etapa, poucos documentos foram anotados e os anotadores praticaram as instruções de anotação e o uso do software *Tagtog*, além de interagirem com os demais discutindo as dúvidas em situações específicas. Cada anotador, rotulou pelo menos duas ocorrências de cada entidade. Ao final, houve revisão das instruções de anotação e da configuração final das entidades a serem rotuladas.

Processo de Anotação do Corpus - Cada documento foi anotado por dois anotadores diferentes. Os anotadores receberam o manual de instruções e, cada um deles teve liberdade para realizar seu trabalho na ordem que preferisse. A primeira etapa do processo de anotação ocorreu em 2,5 meses, quando todos os documentos haviam sido anotados por pelo menos dois anotadores. Depois dessa etapa, a CDJUR-BR entrou no ciclo de avaliação e refinamento.

Avaliação da Concordância entre Anotadores - A Tabela 2 mostra os resultados obtidos do coeficiente Kappa [McHugh 2012] para as categoria de entidades. A categoria Pessoa obteve Kappa 0,79, indicando a maior concordância entre os anotadores. O resultado é considerado "substancial" na interpretação do índice Kappa segundo Landis e Koch [Landis and Koch 1977]. Nesta mesma interpretação, a categoria Prova é considerada "moderada" e foi a que apresentou maior dissenso entre os anotadores.

Tabela 2. Resultados da Concordância entre Anotadores em cada Categoria

Categoria	Anotações	Kappa
Pessoa	15.149	0.79
Prova	1.696	0.46
Pena	205	0.64
Endereço	2.041	0.73
Sentença	106	0.75
Norma	6.216	0.76
Total	25.413	0,69

Processo de Adjudicação - Ao final da fase de anotações, 732 documentos obtiveram coeficiente Kappa médio de 0,69. Para 166 documentos com Kappa inferior a 0,50 foi adotada a estratégia de adjudicação em que um terceiro especialista revisa os casos de desacordo [Hovy and Lavid 2010]. Nestes casos, o revisor recebeu as anotações dos dois anotadores iniciais (união das anotações), e poderia acatar uma das anotações ou mesmo

realizar nova anotação diferente das iniciais. Além disso, no decorrer das anotações, alguns anotadores deixaram os trabalhos, e com isso, 176 documentos remanescentes foram anotados pela equipe de adjudicação.

Refinamento das Normas Legais - Dada a importância da categoria Normas para o domínio jurídico, a equipe de revisores recebeu um relatório com as normas identificadas pelo REN e analisaram se eram, de fato, normas e qual a entidade específica (principal, acessória ou uma jurisprudência). Esta etapa de melhoria e refinamento da CDJUR-BR propiciou a revisão das diretrizes de caráter jurídico destas entidades e dos limites das anotações de normas. Ao final dessa etapa, foram adicionadas 4.338 novas entidades de normas jurídicas. Por fim, a CDJUR-BR contém 1.074 documentos com 44,526 ocorrências de entidades rotuladas, conforme detalhado na Tabela 3.

Tabela 3. Estatística do corpus anotado.

Categoria	Anotações	%	Entidade Nomeada	Anotações	%
Pessoa	24.844	55,80	PES-ADVOG	735	1,65
			PES-AUTOR	1.259	2,83
			PES-AUTORID-POLICIAL	2.012	4,52
			PES-JUIZ	576	1,29
			PES-OUTROS	6.003	13,48
			PES-PROMOTOR-MP	363	0,82
			PES-REU	8.773	19,70
			PES-TESTEMUNHA	2.967	6,66
			PES-VITIMA	2.156	4,84
Prova	3.318	7,45	PROVA	3318	7,45
Pena	407	0,91	PENA	407	0,91
Endereço	2.065	4,64	END-AUTOR	132	0,30
			END-DELITO	466	1,05
			END-OUTROS	355	0,80
			END-REU	693	1,56
			END-TESTEMUNHA	295	0,66
			END-VITIMA	124	0,28
Sentença	172	0,39	SENTENÇA	172	0,39
Norma	13.720	30,81	NOR-ACESSORIA	5.767	12,95
			NOR-JURISPRUDÊNCIA	1.823	4,09
			NOR-PRINCIPAL	6.130	13,77
Total	44.526	100	Total	44.526	100

4. Avaliação Experimental

A avaliação da CDJUR-BR foi guiada pelas Questões de Pesquisa definidas na introdução deste artigo. A *QP1* (Como elaborar uma metodologia de anotações manuais de entidades nomeadas que contemple as especificidades e complexidades do domínio jurídico?) foi respondida nas etapas de desenvolvimento das anotações, descritas na metodologia aqui apresentada, e validada pelas avaliações de concordância entre anotadores (Kappa geral de 0,69). A etapa de refinamento, com uso do REN treinado na CDJUR-BR, numa abordagem *humam-in-the-loop*, possibilitou a adição de 19.113 anotações à CDJUR-BR.

Para responder a *QP2* (A coleção dourada gerada é adequada para o treinamento e validação de modelos de *Legal AI*?), avaliamos a CDJUR-BR na tarefa de REN, através de alguns cenários de experimentos descritos a seguir. Em todos os cenários, os conjuntos de treino, validação e teste ficaram com 68%, 15% e 17% das amostras, respectivamente, observando-se a proporção por categoria no conjunto original.

- **C1. Reconhecimento das entidades específicas da CDJUR-BR**- os dados da CDJUR-BR foram usados para treinar REN específicas. Nosso propósito é avaliar a viabilidade da CDJUR-BR para o treinamento de modelos REN no domínio jurídico brasileiro, em língua portuguesa.

- **C2. Reconhecimento das categorias da CDJUR-BR-** o resultado do REN, treinado com a CDJUR-BR, agrupando as entidades específicas nas seguintes categorias: Pessoa, Legislação, Jurisprudência, e Local. O objetivo é uma análise comparativa com a base LENER-BR.
- **C3. Reconhecimento das entidades da LENER-BR-** neste cenário, os modelos REN foram treinados no conjunto da LENER-BR, em 6 diferentes entidades: Pessoa, Jurisprudência, Tempo, Local, Legislação e Organização.
- **C4. Reconhecimento das entidades da LENER-BR com modelo treinado na CDJUR-BR-** os modelos foram treinados com a CDJUR-BR e testados com os dados do LENER-BR, visando avaliar a capacidade de generalização do modelo REN treinado com a CDJUR-BR quando usado com outros documentos (no caso, os documentos que compuseram a LENER-BR).
- **C5. Reconhecimento das categorias de entidades da CDJUR-BR com modelo treinado na LENER-BR-** os modelos foram treinados com a LENER-BR e testados nas seguintes categorias da CDJUR-BR: Pessoa, Legislação, Jurisprudência e Local. Esse cenário visa avaliar quão capazes serão os modelos treinados com a LENER-BR em reconhecer entidades de outro *corpus* (no caso, os documentos que compuseram a CDJUR-BR).

A Tabela 4 apresenta os resultados obtidos no conjunto de teste para o C1, usando REN desenvolvidos com BI-LSTM + CRF [Graves and Schmidhuber 2005, Hochreiter and Schmidhuber 1997, Lafferty et al. 2001], SPACY [Honnibal and Montani 2017], e BERT [Devlin et al. 2018]. O modelo com o BERT, comparativamente, obteve o melhor desempenho na grande maioria das entidades, alcançando um F1-macro médio = 0,58. Das 21 entidades da CDJUR-BR, 09 (nove) alcançaram F1-Score $\geq 0,70$. Entretanto, 12 entidades tiveram F1-Score inferior a 0,70. Para alguns casos, como END-AUTOR, END-VÍTIMA e SENTENÇA, a causa foi o pequeno número de exemplos anotados para as entidades e os Falsos Positivos (FP) do tipo “O”, os quais representaram mais de 60% dos erros de predição. Para as entidades PES-ADVOG, PES-AUTOR, PES-OUTROS, PES-TESTEMNHA e PES-VÍTIMA, além dos FP com tokens do tipo “O”, também houve uma quantidade equivalente dos erros entre entidades da mesma categoria. Já para a entidade PROVA, a precisão foi excelente (0,87), com poucos FP. No entanto, a cobertura ficou baixa, em 0,33, com FN nas entidades Normas, Prova, Pessoa e, principalmente, nos tokens tipo “O”.

A Tabela 5 apresenta os resultados obtidos para os cenários de C2 a C5. Os experimentos destes cenários foram realizados com o modelo baseado no BERT, visto que este foi a melhor abordagem no cenário C1. Em uma comparação direta de C2 com C3, observamos que os resultados do REN treinado e testado com a CDJUR-BR são inferiores aos resultados do REN treinado e testado com a LENER-BR. A diversidade e quantidade de documentos jurídicos na CDJUR-BR, em relação a quantidade da base LENER-BR contribuiu para este desempenho inferior. Todavia, quando comparamos C4 (modelo treinado com a CDJUR-BR e testado com o LENER-BR) e C5 (modelo treinado com LENER-BR e testado com a CDJUR-BR), verifica-se que a CDJUR-BR obtém um desempenho superior (F1-Macro = 0,68 contra F1-Macro=0,56). Esse resultado indicia que a CDJUR-BR tem maior capacidade de adaptabilidade para reconhecer entidades de outro *corpus* de documentos legais.

Tabela 4. Resultados da Tarefa de REN no Cenário C1, utilizando os modelos BI-LSTM+CRF, SPACY e BERT

Entidade Nomeada	BI-LSTM+CRF	SPACY	BERT	Suporte
END-AUTOR	0.56	0.31	0.33	18
END-DELITO	0.72	0.45	0.73	61
END-OUTROS	0.00	0.02	0.16	81
END-REU	0.55	0.59	0.71	152
END-TESTEMUNHA	0.27	0.26	0.67	68
END-VITIMA	0.06	0.00	0.22	27
NOR-ACESSÓRIA	0.79	0.79	0.82	990
NOR-JURISPRUDÊNCIA	0.90	0.87	0.89	333
NOR-PRINCIPAL	0.67	0.71	0.77	791
PENA	0.56	0.39	0.50	82
PES-ADVOG	0.54	0.22	0.63	122
PES-AUTOR	0.59	0.59	0.56	169
PES-AUTORID-POLICIAL	0.87	0.66	0.90	300
PES-JUIZ	0.79	0.50	0.78	83
PES-OUTROS	0.54	0.44	0.58	1.210
PES-PROMOTOR-MP	0.81	0.27	0.88	57
PES-REU	0.64	0.57	0.71	1.503
PES-TESTEMUNHA	0.57	0.45	0.64	519
PES-VÍTIMA	0.33	0.23	0.46	405
PROVA	0.47	0.29	0.34	461
SENTENÇA	0.00	0.29	0.00	11
F1-micro avg	0.64	0.55	0.67	7.443
F1-macro avg	0.53	0.42	0.58	7.443
F1-weighted avg	0.62	0.54	0.67	7.443

Tabela 5. Resultados de F1-score para o REN na CDJUR-BR e LENER-BR (C2, C3, C4 e C5) utilizando o modelo BERT

Entidade	Cenário de Experimento			
	C2	C3	C4	C5
JURISPRUDÊNCIA	0.89	0.96	0.79	0.48
LEGISLAÇÃO	0.92	0.97	0.92	0.86
LOCAL	0.77	0.77	0.32	0.15
PESSOA	0.83	0.97	0.69	0.76
F1-micro avg	0.85	0.96	0.81	0.60
F1-macro avg	0.85	0.92	0.68	0.56
F1-weighted avg	0.85	0.96	0.79	0.74

5. Conclusão

Neste artigo, apresenta-se a metodologia de anotação manual de documentos jurídicos, que serviu de base para a construção da CDJUR-BR, uma coleção dourada composta de mais de 1000 documentos de processos judiciais brasileiros, contendo 44.526 anotações de 21 entidades específicas das categorias Pessoa, Prova, Pena, Endereço, Sentença e Norma. A avaliação da concordância entre anotadores alcançou medida Kappa de 0,69 para 73% dos documentos, e os demais documentos passaram por revisões por especialistas e etapas de refinamento. Nos experimentos realizados com a CDJUR-BR, os resultados apontaram superioridade do modelo BERT com F1-macro média = 0,58, e testes comparativos entre CDJUR-BR e LENER-BR indicaram que o modelo REN treinado com a CDJUR-BR é superior em reconhecer entidades de outros corpus. Como trabalhos futuros, planeja-se melhorar a desambiguidade entre entidades e realizar mais anotações objetivando reduzir o desbalanceamento entre entidades. Além disso, desenvolver novos modelos REN a fim de selecionar aqueles de melhor desempenho para entidades específicas, realizando comparações com outros *corpus* no domínio jurídico.

Referências

- Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F., Vitória, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., et al. (2022). Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In *International Conference on Computational Processing of the Portuguese Language*, pages 3–14. Springer.
- Angelidis, I., Chalkidis, I., and Koubarakis, M. (2018). Named entity recognition, linking and generation for greek legislation. In *JURIX*, pages 1–10.
- Atdağ, S. and Labatut, V. (2013). A comparison of named entity recognition tools applied to biographical texts. In *2nd International conference on systems and computer science*, pages 228–233. IEEE.
- Cejuela, J. M., McQuilton, P., Ponting, L., Marygold, S. J., Stefancsik, R., Millburn, G. H., Rost, B., Consortium, F., et al. (2014). tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. *Database*, 2014.
- de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Hovy, E. and Lavid, J. (2010). Towards a ‘science’ of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.
- Jiang, R., Banchs, R. E., and Li, H. (2016). Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27.
- Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.
- Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

- Leitner, E., Rehm, G., and Moreno-Schneider, J. (2020). A dataset of german legal documents for named entity recognition. *arXiv preprint arXiv:2003.13016*.
- Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8.
- Santos, D. and Cardoso, N. (2006). A golden resource for named entity recognition in portuguese. In *International Workshop on Computational Processing of the Portuguese Language*, pages 69–79. Springer.
- Schmitt, X., Kubler, S., Robert, J., Papadakis, M., and LeTraon, Y. (2019). A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343. IEEE.
- Silva, R. L. d., Hoch, P. A., and Righi, L. M. (2013). Transparência pública e a atuação normativa do cnj. *Revista direito GV*, 9:489–514.
- Yadav, V. and Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

A call for a research agenda on fair NLP for Portuguese

Luiz Fernando F. P. de Lima¹, Renata Mendes de Araujo^{2,3,4}

¹ Centro de Estudos e Sistemas Avançados do Recife (CESAR) - Recife, PE, Brasil

²Universidade Presbiteriana Mackenzie - São Paulo, SP, Brasil

³Universidade de São Paulo (USP) - São Paulo, SP, Brasil

⁴Escola Nacional de Administração Pública (ENAP) - Brasília, DF, Brasil

lffpl@cesar.org.br, renata.araujo@mackenzie.br

Abstract. *Diverse areas widely apply artificial intelligence and natural language processing (NLP) tools to their contexts. However, these algorithms present ethical issues, such as biased and discriminatory decisions. For example, representation biases in NLP can result in discriminatory behavior towards race and gender. Works have been addressing this issue and seeking to build fair NLP solutions, however they mainly focus on Anglo-Saxon languages. This work aims to challenge the scientific community in order to stimulate and motivate further research in the fair NLP specifically for the Portuguese language. To achieve this, a literature review was conducted to identify existing research efforts and indicate future directions.*

1. Introduction

Different biases, such as historical biases, representation biases, evaluation biases, and human interpretation biases, can be embedded into a machine learning (ML) model during its training [Ruback et al. 2021]. The incorporation of these biases by the model can result in unfair and discriminatory outcomes.

We can also observe cases of algorithmic discrimination in NLP tools. For example, Amazon’s algorithm for recruiting and selecting employees that penalized resumes belonging to women and identified the applicant’s gender, even if this information was omitted [Dastin 2018]. On the other hand, recent literature presents attempts to achieve fair NLP algorithms for different tasks through various techniques such as data augmentation, gender marking and learning gender-neutral embeddings [Mehrabi et al. 2021, Bolukbasi et al. 2016, Leavy 2018].

However, these fair NLP researches mostly use English databases, not worrying about whether their approaches work when concerning other languages. In this scenario, we can raise concerns about the democratization of these fair NLP solutions for different languages, especially for the Portuguese language. Therefore, it is essential to stimulate and advance research in the aspect of fair NLP for Portuguese, making this technology fairer for a larger segment of society [Camões - Instituto da Cooperação e da Língua 2023].

In this sense, the main goal of this work is to identify research gaps, and outline a research agenda in order to motivate scholarly peers to foster a broader

democratization of fair NLP solutions for Portuguese. As a first step to achieve this goal we focused on mapping, through a literature review, fair NLP solutions that are concerned with solving problems of algorithmic discrimination in Portuguese.

2. Background

In this work, we aim to outline a research agenda for fair NLP concerning with Portuguese. We understand fair NLP as techniques that attempts to mitigate representation biases which can lead algorithms to discriminatory behavior, such as, denigration, stereotypes, recognition, and under-representation [Sun et al. 2019].

Fair NLP works also points out and discuss the societal and ethics implications of these problems and introducing metrics, evaluations, and fair architectures/models for diverse NLP tasks. For instance, [Bolukbasi et al. 2016] investigated the gender bias inherent in word embeddings and proposed a framework to mitigate such behavior. Moreover, [Bender et al. 2021] presented some of the social impacts, limitations, and potential harms associated with language models.

The literature present some surveys and systematic reviews that map research from the perspective of fair NLP. For example, [Mehrabi et al. 2021] bring a broad survey on algorithmic fairness, including mapping fair NLP solutions. However, the authors do not present their research methodology, leaving aside the steps for reproducing the performed review. Given the many references raised, this work can help identify and aggregate research not found by our own methodology.

Among the research presented in the survey, we highlight the approach proposed by [Font and Costa-Jussa 2019], which focuses on mitigating gender biases in translating sentences between English and Spanish. In addition, from the perspective of mitigating gender biases in machine translation (MT) tasks, the work by [Vanmassenhove et al. 2018] uses a multilingual dataset with examples of texts in all languages present in the European Union, including Portuguese.

In their literature review, [Sun et al. 2019] categorize the research under four perspectives of biases: denigration, stereotypes, recognition, and under-representation. Although this work does not present any methodological aspect to map the presented papers, its final section reinforce our concerns on building a research agenda focusing on the mitigation of biases in languages other than English.

Finally, [Blodgett et al. 2020] present a survey with a more critical perspective on NLP research that address issues of bias, mainly due to vagueness or inconsistency in the motivations of the reviewed papers.

3. Method

The methodology applied to this literature review is based on the procedure presented by [Blodgett et al. 2020], however adding guarantees that we are returning research papers that consider Portuguese. The steps and results are detailed below.

We seek to answer the following research questions: (1) What papers are concerned with examining NLP techniques just for Portuguese? (2) What biases do the literature approaches aim to mitigate? And for which NLP tasks? (3) Are there

database resources in Portuguese that can be used in the context of fair NLP? (4) What is the maturity of reproducibility and openness of information of these works?

To select the articles, we used research strings with terms commonly present in studies on NLP, adding constraints to properly select works concerning Portuguese: ((“bias” AND (“fairness” OR “equity”)) AND (“natural language processing” OR “nlp”) AND “portuguese”) OR ((“viés” AND (“justiça” OR “equidade”)) AND (“processamento de linguagem natural” OR “NLP”) AND “português”).

The IEEE Xplore, ScienceDirect, and Periódicos CAPES were used as research sources, collecting research works in the five-year interval (2018 - 2022). In addition, as a guarantee of not excluding relevant work, we manually searched for papers in some relevant conferences in the areas of AI/ML, NLP, and ethics in AI: ICML, NeurIPS, AIES, FAccT, WWW, BRACIS, ENIAC, PROPOR, and STIL.

Using the search string in research sources, no papers were returned. With the manual search in the conferences’ proceedings, we could identify two papers that fit this review. Finally, we add to this set the work of [Vanmassenhove et al. 2018], pointed in Section 2 as a potential reference for the analysis. As defined by [Blodgett et al. 2020], the scope of the papers was analyzed as an inclusion criterion. Those articles that talk about NLP only with text are included, thus omitting other works that use speech. In this last step, we kept the raised three articles.

4. Results

Table 1 presents an overview of the analysis, pointing out the answers to the research questions. All papers follow the positivist research paradigm, bringing algorithmic proposals and experiments to analyze and mitigate gender biases in NLP. [Vanmassenhove et al. 2018, Cho et al. 2021] focus on MT tasks, while [Santana et al. 2018] work on de-biasing word embeddings for analogy tasks. In addition, we observed how the quality of the works in relation to issues of reproducibility of their experiments vary due the presentation of reproducibility details.

Table 1. Overview of the analyzed papers concerning the research questions

Reference	Getting gender right in neural machine translation [Vanmassenhove et al. 2019]	Is there gender bias and stereotype in portuguese word embeddings? [Santana et al. 2018]	Towards cross-lingual generalization of translation gender bias [Cho et al. 2021]
Source	Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018	International Conference on the Computational Processing of Portuguese (PROPOR) 2018	ACM Conference on Fairness, Accountability, and Transparency 2021
NLP Task	Machine translation	Word embeddings analogies	Machine translation
Studied Bias	Gender bias	Gender bias	Gender bias
Dataset	Corpus built from Europarl	Corpus proposed by [Hartmann et al. 2017]	Corpus built from the proposed by [Cho et al. 2019] + modifications with systematic process for translation
Reproducibility	Moderated	High (with limitations)	High

Considering the differences in language structures in terms of gender and observing the loss of information in automated translations, [Vanmassenhove et al. 2018] seek to mitigate gender biases in MT that result in morphologically incorrect translations. Their propose focuses on the use of gender information to help MT algorithms to perform translations with better quality. The assessments for translations into FR and PT suggest that the approach has the potential to improve gender agreement in translations between EN-PT.

[Santana et al. 2018] aim to analyze and remove gender biases in Portuguese word embeddings for the analogy task. The authors propose an evaluation pipeline in three stages: 1) use of the word2vec model with strategy proposed

by [Hartmann et al. 2017]; 2) application of the debiasing algorithm proposed by [Bolukbasi et al. 2016] and; 3) model accuracy assessment. The authors focus on mitigating bias in Portuguese word embeddings and investigating its effects on the accuracy of the model. However, they do not propose any adjustments to the applied debiasing algorithm to account for the specificities of Portuguese gender structures.

[Cho et al. 2021] also present a approach based on algorithmic advances to evaluate gender biases in machine translation tools. However, they aim to conceptualize and consider several linguistic aspects, such as the presence of gender-neutral pronouns, agreement of articles concerning gender, and derivation of the noun according to its gender, on the languages presented, to build a more generic and agnostic model to the language pairs for this assessment. Those concerns also regard to the gender structures present in the Portuguese language.

5. Outlining a Research Agenda

In this current work, we could not find any work in the chosen search databases, only two papers were identified in conference proceedings, and one more research was added to this review because it was referenced in a previously analyzed paper.

These results indicates how many perspectives still open to be explored in research on fair NLP in Portuguese. For example, future works can continue to examine the interchangeability of existing fair NLP solutions that mitigate gender bias for Portuguese, e.g., one could evaluate the use of other techniques such as those presented by [Mehrabi et al. 2021].

Other opportunities are related to the concern with the regional linguistic variations of Portuguese. Similarly, concerning in mitigating harms to specific and marginalized groups, additional research efforts may focus on exploring linguistic variations within a particular country, such as Brazil [Drager et al. 2021, Guy 1981]. Furthermore, works could extend to variations and dialects specific to marginalized groups, such as the LGBTQIA+ community.

Beyond gender biases, future work needs to bring efforts to the mitigation of racial biases in NLP in Portuguese. The racial perspective has been explored in the context of fair NLP by the international community [Blodgett and O'Connor 2017] and needs to be amplified for the Portuguese context.

For all these perspectives, it is essential to engage in the endeavor of seeking more diverse and representative datasets. Research efforts should focus on examining the presence of biases and sources of harm within already published datasets. Additionally, it is crucial to make a concerted effort to evaluate datasets when a group or entity intends to construct and publish a new dataset. In this regard, one can adopt the methodology proposed by [Geburu et al. 2021].

Finally, we point out opportunities for carrying out research using other paradigms. As suggested by [Araujo et al. 2017] for the IS area, research in NLP, can use interpretive approaches in order to present contextualized works to society, analyzing the impacts of proposed solutions to the world. Likewise, critical research is welcome in this area and may refer to, for example, critical works on algorithmic racism in facial recognition tools [Buolamwini and Geburu 2018, Silva 2020].

References

- Araujo, R., Fornazin, M., and Pimentel, M. (2017). Uma análise sobre a produção de conhecimento científico nas pesquisas publicadas nos primeiros 10 anos da isys (2008-2017). *iSys-Brazilian Journal of Information Systems*, 10(4):45–65.
- 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, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of” bias” in nlp. *arXiv preprint arXiv:2005.14050*.
- Blodgett, S. L. and O’Connor, B. (2017). Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.
- Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.
- Camões - Instituto da Cooperação e da Língua (2023). Dados sobre a língua portuguesa. https://www.instituto-camoes.pt/images/img_agenda2023/Dados_sobre_a_l%C3%ADngua_portuguesa_2023.pdf.
- Cho, W. I., Kim, J., Yang, J., and Kim, N. S. (2021). Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 449–457.
- Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.
- Drager, K., Rilliard, A. O. B., Vieira, M. d. S. M., and Wiedemer, M. L. (2021). Linguistic varieties in brazil and beyond. *Revista Diadorim*, 23(1):24–33.
- Font, J. E. and Costa-Jussa, M. R. (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.
- Guy, G. R. (1981). Linguistic variation in brazilian portuguese: Aspects of the phonology, syntax, and language history.

- Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.
- Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.
- Ruback, L., Avila, S., and Cantero, L. (2021). Vieses no aprendizado de máquina e suas implicações sociais: Um estudo de caso no reconhecimento facial. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, pages 90–101, Porto Alegre, RS, Brasil. SBC.
- Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there gender bias and stereotype in portuguese word embeddings? *arXiv preprint arXiv:1810.04528*.
- Silva, T. (2020). Visão computacional e racismo algorítmico: branquitude e opacidade no aprendizado de máquina. *Revista ABPN*, 12:428–448.
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.
- Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Avaliação do senso comum em modelos de linguagem através de benchmarks: Desafio de Winograd aplicado ao ChatGPT em português brasileiro

Thiago Gomes do Nascimento, Diogo Cortiz

Pontifícia Universidade Católica de São Paulo (PUC– SP) – São Paulo, SP – Brasil

thiago.gnascimento1@gmail.com, diogocortiz@gmail.com

Abstract. *The assessment of language models with benchmarks is presented as an effective way of evaluating their comprehension limits. In this regard, the Winograd Schema Challenge, which aims to assess common sense through pronoun disambiguation tasks, has led to the development of different metrics and datasets. When applying a translation of the Winograd Challenge to Brazilian Portuguese to ChatGPT, we identified comparable results to those obtained in English. However, these results must be analyzed with caution, considering the potential biases in the model training process and the existing gaps in the reasoning dimensions covered by the available evaluation methods.*

Resumo. *O desempenho em benchmarks é apresentado como uma forma de avaliação efetiva dos limites de compreensão dos modelos de linguagem. Neste sentido, o desafio de esquemas de Winograd, que se propõe a avaliar o senso comum por meio de tarefas de desambiguação de pronomes, deu origem a diferentes métricas e datasets. Ao aplicar a tradução do desafio de Winograd ao ChatGPT em português brasileiro, identificamos resultados equiparáveis aos obtidos em inglês. Contudo, é preciso ter cautela ao interpretar estes dados, visto que existem vieses associados ao treinamento dos modelos e lacunas quanto às dimensões de raciocínio contempladas pelos métodos de avaliação disponíveis.*

1. Introdução

Os modelos generativos de linguagem, como o GPT desenvolvido pela Open AI, o LaMDA, presente no Google Bard, e o LLaMa da Meta têm ganhado notoriedade pela capacidade de processamento de textos com resultados semelhantes aos produzidos por humanos. Apesar do crescimento das bases de dados e do aprimoramento dos algoritmos, estes modelos estão associados a aspectos sintáticos [Floridi, 2023]. Devido à característica de combinar sequências linguísticas de acordo com cálculos estatísticos, sem um entendimento semântico, são classificados como papagaios estocásticos [Bender et al., 2021].

Turing [1950] foi pioneiro na criação de uma definição operacional de inteligência. Diante da necessidade latente de avaliar os limites de compreensão dos modelos, tanto entre ferramentas quanto em relação ao ser humano, o Desafio de Winograd [Levesque et al., 2012] surgiu como uma evolução do teste de Turing, no intuito de avaliar o senso comum, habilidade intrinsecamente antropomórfica, de

maneira objetiva. O desafio consiste em responder uma pergunta binária associada a uma frase ambígua, como por exemplo: a medalha não cabe na mala porque ela é muito grande. O que é muito grande? a) a mala ou b) a medalha?

Apesar dos significativos avanços em termos de processamento, a predominância no inglês na construção e no treinamento dos modelos desperta questões relativas à confiabilidade dos resultados fornecidos para idiomas sub-representados [Petrov et al., 2023]. Partindo da tradução do Desafio de Winograd para o português brasileiro [Melo et al., 2019], avaliamos o desempenho do ChatGPT na tarefa de desambiguação de pronomes a fim de identificar se haveria alguma disparidade com os resultados publicados em inglês.

2. Winograd e a evolução dos benchmarks

O jogo da imitação [Turing, 1950] foi proposto como uma maneira de avaliar a inteligência da máquina. Através da interação por perguntas, uma pessoa exercendo o papel de juiz deve identificar qual dos respondentes é um humano e qual é uma máquina. Uma das críticas ao teste consiste na objeção da Lady Lovelace, segundo a qual a máquina faz apenas aquilo que mandamos. O argumento de Turing justifica que, apesar de receber instruções básicas, um computador capaz de aprender de forma autônoma poderia fazer o que não foi determinado anteriormente por seu programador [French, 2000].

A falta de objetividade foi considerada com outra deficiência do teste de Turing, que dependia da avaliação subjetiva do julgador humano. Utilizando a contribuição de Winograd [1972] no estudo da compreensão da linguagem natural, Levesque et al. [2012] propuseram o desafio de esquemas de Winograd. O teste consiste na avaliação de um grupo de declarações com sentido dúbio, cuja resposta não pode ser obtida na frase, apenas através de conhecimento prévio, ou seja, o senso comum que permite a associação semântica.

Devido à limitação de serem elaborados inicialmente em inglês, os esquemas foram traduzidos para diferentes idiomas como português [Melo et al., 2019], francês [Amsili; Seminck, 2017], húngaro [Vadász; Ligeti-Nagy, 2022], mandarim [Bernard; Han, 2020] e russo [Shavrina et al., 2020]. Além disso, diversos autores apresentaram evoluções dos esquemas. Construído com a proposta de ser um dataset multilíngue, o Wino-X [Emelin; Sennrich, 2021] contempla esquemas em alemão, francês e russo alinhados com as versões originais em inglês.

Além disso, diversos autores apresentaram evoluções das bases de dados utilizadas, como o Winogrande [Sakaguchi et al., 2021], que contempla 44.000 esquemas. Algumas alternativas encontradas para a ampliação do volume de esquemas são o Winoflexi [Isaak; Michael, 2019], que utiliza *crowdsourcing* para o desenvolvimento de novas sentenças e o Wininventor [Nicos; Michael, 2020] que busca automatizar a criação de esquemas. No Winologic [He et al., 2021] novas frases foram construídas utilizando teoremas lógicos.

Na adaptação WNLI [Wang et al., 2018], o desafio de Winograd foi reformulado como uma tarefa de inferência de linguagem natural. Neste caso, o formato das tarefas é composto por três partes. Premissa: a medalha não cabe na mala porque ela é muito grande. Hipótese: a medalha é muito grande. Resposta: verdadeiro/falso. Esta versão foi

adicionada ao benchmark GLUE, em conjunto com outras tarefas. O aprimoramento trazido pelo SUPERGLUE [Wang et al., 2019], apresentado como detentor de um maior nível de dificuldade, considera a versão original dos esquemas. Storks et al. [2019] apresentam uma classificação para os benchmarks que realizam a avaliação de raciocínio de senso comum para compreensão de linguagem natural, de acordo com o tipo de atividade testada. Um levantamento mais recente, realizado por Davis [2023] lista 139 benchmarks, sendo 102 de texto, 18 para imagens, 12 de vídeos e 7 de ambientes físicos.

3. Método

Dos 285 esquemas do desafio de Winograd, a tradução elaborada por Melo et al. [2019] classifica 8 como sem equivalência no português. Assim, para este estudo foram utilizados 277 esquemas. A base contempla a frase principal de cada etapa do teste seguida por colunas que indicam o trecho ambíguo e o pronome que é necessário para identificar a referência. Além disso, também estão descritas quais são as alternativas e a resposta correta para cada tarefa.

"schema": "Os vereadores recusaram a autorização aos manifestantes porque eles temiam a violência.",
"snippet": "eles temiam a violência.",
"pronoun": "eles",
"correct_answer": "A",
"substitution_a": "Os vereadores",
"substitution_b": "Os manifestantes"

Neste experimento, foi utilizada a versão grátis do ChatGPT, através da interface disponível no endereço <https://chat.openai.com/>. O *prompt* utilizado seguiu o padrão original de formulação dos esquemas, exposto na seção 1. As tarefas foram adicionadas uma a uma, iniciando pela frase principal do desafio. Em seguida, era perguntado quem ou que estava associado ao trecho ambíguo, sempre com duas opções de resposta, identificadas como “a” e “b”, conforme é possível observar na figura 1. As alternativas retornadas pelo ChatGPT foram coletadas e adicionadas a uma nova coluna da base. Comparando a coluna das respostas corretas com a coluna das respostas obtidas foi possível calcular o percentual de acerto em relação ao total de perguntas realizadas.

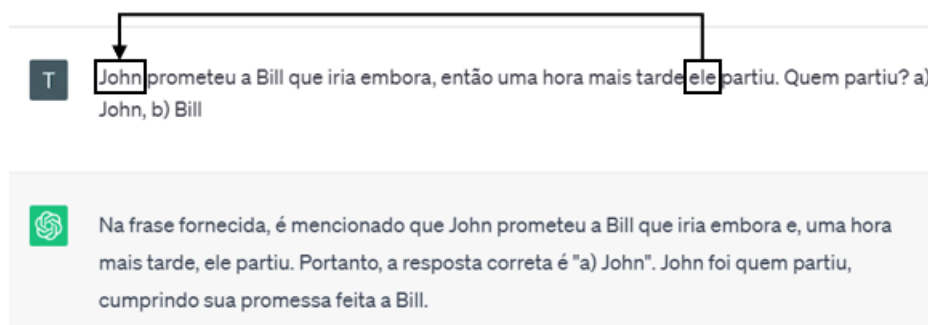


Figura 1. Exemplo de interação com o ChatGPT para obtenção das respostas.

4. Resultados

O ChatGPT, que incorpora o modelo GPT-3.5 apresentou um percentual de acerto de 87,5% em relação às 277 perguntas do desafio de Winograd respondidas em português brasileiro. Na tabela 1, também é possível observar que o resultado obtido é similar ao apresentado por modelos de linguagem de grande porte em tarefas de desambiguação de pronomes no idioma inglês.

Quanto às versões do teste, WSC285 refere-se à lista completa do desafio, que contém 12 tarefas além dos 273 esquemas presentes na versão anterior, conhecida como WSC273. O presente estudo teve uma diferença de apenas 0,8% abaixo do modelo GPT-3 em inglês [Brow et al., 2020]. Ao contrapor os resultados com o GPT 3.5, é preciso considerar que os dados divulgados são referentes à versão Winogrande do desafio, que conforme mencionado anteriormente, possui uma base mais extensa. Na comparação com o RoBERTa [Sakaguchi et al., 2021], evolução do modelo BERT com melhorias de arquitetura, o ChatGPT apresentou uma diferença de 2,6% abaixo do resultado para a versão WSC 273.

Tabela 1. Resultado obtido comparado com desempenho registrado em estudos anteriores.

Modelo	Idioma	Desempenho	Versão do teste	Fonte
GPT-3.5	Português	87,5%	WSC 285	Própria
GPT-3	Inglês	88,3%	WSC 273	Brown et al. [2020]
GPT-3.5	Inglês	81,6%	Winogrande	OpenAI [2023]
RoBERTa	Inglês	90,1%	WSC 273	Sakaguchi et al. [2021]

5. Conclusão

Apesar da pontuação da apresentada, não é possível concluir que o ChatGPT possui senso comum nem que o desempenho do modelo em português para outras tarefas é tão confiável quanto a versão original, em inglês. O desafio de Winograd está disponível na internet desde 2012, sendo provável que tenha feito parte dos dados de treinamento dos atuais modelos de linguagem. Além disso, o sucesso em uma tarefa específica não é uma métrica confiável para avaliar o senso comum.

Dentre estas tarefas, a desambiguação de pronomes contempla apenas uma pequena parcela do senso comum que é preciso para compreensão da linguagem. Considerando os demais *benchmarks* disponíveis para senso comum, apenas duas dimensões de raciocínio são avaliadas adequadamente: o taxonômico, que se refere a classificações e ao conhecimento enciclopédico, e o numérico, que abrange cálculos e quantidades. As demais dimensões, do raciocínio (temporal, psicológico, espacial, físico, biológico, social, comparativo, meta-raciocínio) não são contempladas ou são abordadas parcialmente. Assim, a construção de novos *benchmarks* deve considerar aspectos ainda não explorados do senso comum, a fim de propiciar uma avaliação mais adequada da capacidade dos modelos [Kocijan et al., 2023; Davis, 2023].

Referências

- Amsili, P.; Seminck, O. (2017) “A Google-Proof Collection of French Winograd Schemas”, Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), p. 24-29.
- Bender, E. M. et al. (2021) “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, p. 610-623.
- Bernard, T.; Han, T. (2020) “Mandarinograd: A Chinese Collection of Winograd Schemas”, Proceedings of the Twelfth Language Resources and Evaluation Conference, p. 21-26.
- Brown, T. B. et al. (2020) “Language models are few-shot learners”, NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, p. 1877-1901.
- Davis, E. (2023) “Benchmarks for Automated Commonsense Reasoning: A Survey”, arXiv:2302.04752v2, <https://doi.org/10.48550/arXiv.2302.04752>
- Emelin, D.; Sennrich, R. (2021) “Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution”, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, p. 8517-8532.
- Floridi, L. (2023) “AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models”, Philosophy & Technology 36 (15), p. 1-7.
- French, R. M. (2000) “The turing test: The first 50 years”, Trends in Cognitive Sciences 4 (3), p. 115-122.
- He, W. et al. (2021) “WINOLOGIC: A Zero-Shot Logic-based Diagnostic Dataset for Winograd Schema Challenge.”, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, p. 3779–3789.
- Isaak, N.; Michael, L. (2019) “WinoFlexi: A Crowdsourcing Platform for the Development of Winograd Schemas” In: Liu, J., Bailey, J. (eds) AI 2019: Advances in Artificial Intelligence. AI 2019. Lecture Notes in Computer Science 11919.
- Kocijan, V. et al. (2023) “The Defeat of the Winograd Schema Challenge”, arXiv:2201.02387v3, <https://doi.org/10.48550/arXiv.2201.02387>
- Levesque, H. J.; Davis, E.; Morgenstern, L. (2012) “The Winograd Schema Challenge”, Thirteenth international conference on the principles of knowledge representation and reasoning.
- Melo, G. S. D.; Imaizumi, V. A.; Cozman, F. G. (2019), “Winograd Schemas in Portuguese”, Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2019), p. 787–798.
- Nicos, I.; Michael, L. (2020) “Wininventor: A Machine-driven Approach for the Development of Winograd Schemas”, Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, p. 26-35.

- OpenAI (2023) “GPT-4 Technical Report”, arXiv:2303.08774v3, <https://doi.org/10.48550/arXiv.2303.08774>
- Petrov, A. et al. (2023), “Language Model Tokenizers Introduce Unfairness Between Languages”, oarXiv:2305.15425v1, <https://doi.org/10.48550/arXiv.2305.15425>
- Pires, R. et al. (2023), “Sabiá, Portuguese Large Language Models”, arXiv:2304.07880v2, <https://doi.org/10.48550/arXiv.2304.07880>
- Sakaguchi et al. (2021), “WinoGrande: An Adversarial Winograd Schema Challenge at Scale”, *Communications of the ACM* 64(9), p. 99-106.
- Shavrina, T. et al. (2020), “RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark”, *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, p. 4717-4726.
- Storks, S.; Gao, Q.; Chai, J. Y. (2019) “Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches”, arXiv:1904.01172v3, <https://doi.org/10.48550/arXiv.1904.01172>
- Turing, A. M. (1950) “Computing machinery and intelligence”, *Mind* LIX (236), p. 433-460.
- Vadász, N.; Ligeti-Nagy, N. “Winograd schemata and other datasets for anaphora resolution in Hungarian”, *Acta Linguistica Academica* 69 (4), p. 564-580.
- Wang, A. et al. (2018) “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, p. 353-355.
- Wang, A. et al. (2019) “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, p. 3266-3280.
- Winograd, T. (1972) “Understanding natural language”, *Cognitive Psychology* 3(1), p. 1 – 191.

A Sentiment Analysis Benchmark for Automated Machine Learning Applications and a Proof of Concept in Hate Speech Detection

Marília Costa Rosendo Silva, Vitor Augusto de Oliveira,
Thiago Alexandre Salgueiro Pardo

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

marilia.costa.silva@usp.br,
vitor.augusto.oliveira@alumni.usp.br,
taspardo@icmc.usp.br

Abstract. *Automated Machine Learning (AutoML) is a relevant research endeavor as it allows for speeding up and easing the development of new applied solutions using Artificial Intelligence. This paper addresses the challenge of providing standardized datasets for sentiment analysis in English and proposes an AutoML benchmark, resulting in 46 preprocessed datasets. More than this, a proof of concept is carried out for the hate speech detection task to present the potentialities of the proposed benchmark.*

Resumo. *O Aprendizado de Máquina Automático (AutoML) é uma área de pesquisa relevante, pois permite acelerar e facilitar o desenvolvimento de novas soluções aplicadas usando Inteligência Artificial. Este artigo aborda o desafio de fornecer conjuntos de dados padronizados para análise de sentimentos em inglês e propõe um benchmark de AutoML, resultando em 46 conjuntos de dados pré-processados. É realizada uma prova de conceito para a tarefa de detecção de discurso de ódio para apresentar as potencialidades do benchmark proposto.*

1. Introduction

Natural Language Processing (NLP) aims at enabling machines to deal with human languages. The tasks of Sentiment Analysis (SA) are among the most useful and challenging ones, with interest of academic, commercial, and government areas.

In SA research, Machine Learning (ML) techniques have been the dominant approach. Developing an ML solution, however, can be complex for non-experts. For this reason, Automated Machine Learning (AutoML) has gained importance, providing resources to speed up tuning and making ML approaches more accessible [Guyon et al. 2016]. There are a few dozen available AutoML frameworks/systems, and a system that performs well on some tasks may have a lower performance on others [Škrlj et al. 2021]. Therefore, standardized comparison practices, such as benchmarks, can contribute to the traceability of the literature.

In this context, we explore AutoML for SA tasks. This work brings two core contributions: it furnishes 46 preprocessed datasets for different SA tasks; and, as Proof of Concept (PoC), some experiments with statistical evaluation to support the empirical findings comprising hate speech detection datasets and AutoML Systems.

2. Related Works

There are several initiatives on AutoML and on benchmarking some areas and tasks, but there are limited efforts focused on SA. [Blohm et al. 2021] used 13 text datasets for classification tasks, including polarity classification, with only a general evaluation. RAFT [Alex et al. 2021] is a Few-Shot Learning benchmark and uses news articles, domain-specific datasets, one Hate Speech Dataset, and another with complaints on Twitter.

Regarding comparative evaluations and statistical tests, there are several approaches in the literature and, not rarely, limited understanding of the appropriate metrics. [Demšar 2006] recommended the non-parametric Friedman test when assessing multiple classifiers in multiple datasets, and the post-hoc Nemenyi test to assess pairwise differences when the null hypothesis is rejected.

3. Dataset Collection and Preprocessing

To produce a benchmark for the SA area, it is necessary to collect and preprocess datasets, for later selecting and applying AutoML techniques, and standardizing experiment setups. Part of the procedures was based on [Pineau et al. 2020].

The data sources included UC Irvine, GitHub, Hugging Face, Kaggle, SemEval, TensorFlow, OpenML, and research articles. Our work only considered datasets without synthetically generated instances. The authors managed to split all the datasets into two non-overlapping sets (training and test sets).

The fields with text data and the one with the target feature were renamed as “text” and “label”, respectively. This standardized denomination facilitates large-scale experiments. Moreover, the preprocessing steps were customized for each dataset. In addition, the text could have more than one language. Nevertheless, this work addressed data exclusively in English that was identified with the use of fastText [Bojanowski et al. 2017]. Instances that combined English and another language were kept and Regular Expressions were used to remove Cyrillic, Chinese, or Arabic characters.

After the preprocessing steps, instances smaller than three characters or that were duplicates were excluded. Appendix A lists all the datasets and their corresponding tasks. The adopted preprocessing steps do not harm the performance of most NLP tasks.

4. AutoML Experiments

For our experiments, the task of Hate Speech Detection was chosen. It brings all the challenging characteristics of SA tasks (as subjectivity, different writing styles, and high dependence on text genre and domain) and represents a severe disease in our modern society. Handling such scientific and social problem is of utmost importance.

Thirteen preprocessed Hate Speech Detection datasets were used (see Appendix A for the citations). They were selected due to technical considerations: 10 datasets were binary, and they had different and contrasting degrees of class imbalance.

Three AutoML systems and one classifier were adopted. These systems were selected based on previous works and their underlying assumptions. The foundations of AutoGluon [Erickson et al. 2020], Auto-Sklearn [Feurer et al. 2015], and TPOT [Olson and Moore 2016] are ensembles, meta-learning, and genetic programming, respectively. The classifier was Logistic Regression (LR) combined with Random Search (RS). All the approaches had the same conditions: wall time of 15 minutes per dataset, Python Version 3.9.13, Ubuntu 22.04.2, and up to 4GB RAM.

After preprocessing all the hate speech datasets, this work embedded the text with

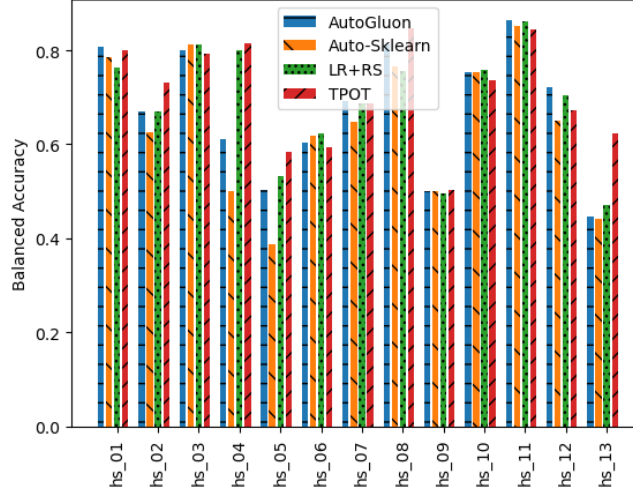


Figure 1. Bar Chart with Balanced Accuracy per Dataset

	Friedman		Kruskal-Wallis	
	χ_{n-1}	$p - value$	H	$p - value$
Values	5.0859	0.1656	1.4216	0.7005

Table 1. Statistics of the Non-Parametric Tests

Sentence BERT [Reimers and Gurevych 2019]¹. Figure 1 presents the bar charts corresponding to the AutoML systems and LR and their balanced accuracy per dataset.

On average, the best AutoML method is TPOT, with a mean of 71.01% of balanced accuracy. Auto-Sklearn has the largest standard deviation (14.97%). TPOT has the highest difference between the mean and median (2.24%), followed by Logistic Regression (1.82%), AutoGluon (1.60%), and Auto-Sklearn (0.61%). The maximum balanced accuracy is 86.50% (AutoGluon), whereas the smallest is 38.64% (Auto-Sklearn).

The results were evaluated with non-parametric hypothesis testing (Friedman and Kruskal-Wallis Tests). The statistics and p-values are presented in Table 1. With $\alpha < 0.05$, it is not possible to reject the null hypothesis, which means that there is no suggestion of statistically significant difference among the systems. This is a very interesting finding as it shows that different AutoML systems may prove to have similar results.

It is also interesting to evaluate the potentiality of the AutoML approaches when compared to the original results produced for the datasets. Unfortunately, it was not viable to perform comparisons for all the datasets (e.g., some of them did not have train-test splits by default). However, for three datasets, it was possible to supply fair comparisons.

Tables 2, 3 and 4 display performance metrics for three different datasets (*hs_03* [de Gibert et al. 2018], *hs_04* [Jigsaw 2018], and *hs_07* [Zampieri et al. 2019]) using AutoML systems and manual hyperparameter tuning from related literature (the evaluation metrics are the ones of the corresponding related works). The tables compare these methods based on various metrics, using different criteria for fair comparisons. Notably, there is a discernible difference between the best results from AutoML and the reported literature, with the latter performing about 9.3% better on average. However, considering that AutoML offers a more general solution and does not require specific tuning, it proves

¹"sentence-transformers/all-MiniLM-L6-v2"

beneficial by relieving users of the task of building ML solutions from scratch. Despite the performance gap, the results achieved by AutoML are deemed satisfactory, suggesting that further investment in this approach is worthwhile.

Dataset	AutoGluon	Auto-Sklearn	LR+RS	TPOT	[de Gibert et al. 2018]
hs_03	0.817	0.798	0.819	0.803	0.892

Table 2. F1-Score for Benchmarking hs_03 [de Gibert et al. 2018]

Dataset	AutoGluon	Auto-Sklearn	LR+RS	TPOT	[Jigsaw 2018]
hs_04	0.9229	0.9060	0.9278	0.8572	0.9885

Table 3. Accuracy for Benchmarking hs_04 [Jigsaw 2018]

Dataset	AutoGluon	Auto-Sklearn	LR+RS	TPOT	[Zampieri et al. 2019]
hs_07	0.67	0.67	0.71	0.70	0.80

Table 4. F1-Score Macro for Benchmarking hs_07 [Zampieri et al. 2019]

5. Final Remarks

This paper introduces a benchmark dataset for SA, performing a PoC on detecting hate speech, showing that AutoML is a challenge that is worthy to follow. Overall, 46 pre-processed datasets are proposed. To the best of our knowledge, this is the first work that accomplishes this.

The proposed benchmark can be expanded with new datasets. They should have the same rationale – train-test splits, single class per instance, the same classes in the training and test sets, using the same codification, providing Python implementation with all the preprocessing steps (e.g., regular expressions and sorting functions), among others – and an available BibTex to furnish use in academia and by practitioners. These criteria can ensure sustainable growth and an update of the proposed benchmark.

Some limitations of this work are that it comprised only English datasets and that some datasets requiring credentials (e.g., using Twitter API to retrieve posts based on identifiers) might lose instances due to social media policy violations. Future research opportunities include improving algorithm initialization and evaluating other classification strategies.

A. Datasets

The next paragraph summarizes the tasks and datasets. They are split into six SA tasks².

Emotion Detection (*ed*): [Strapparava and Mihalcea 2007], [Saravia et al. 2018], [Demszky et al. 2020], [Chakravarthi 2020], [Sosea et al. 2022]; **Fake News Detection** (*fn*): [Wang 2017], [Pérez-Rosas et al. 2018], [Torabi Asr and Taboada 2018], [Torabi Asr and Taboada 2018], [Thorne et al. 2018], [Abu Salem et al. 2019], [Thorne et al. 2019], [Shahi and Nandini 2020], [Weinzierl and Harabagiu 2022], [Weinzierl and Harabagiu 2022]; **Hate Speech Detection** (*hs*): [Waseem and Hovy 2016], [Davidson et al. 2017], [de Gibert et al. 2018], [Jigsaw 2018], [Founta et al. 2018], [Basile et al. 2019], [Zampieri et al. 2019], [Hugging Face 2019], [Gautam et al. 2020], [Mollas et al. 2020], [Grosz and Conde-Cespedes 2020], [Kaggle 2020c], [Mathew et al. 2021]; **Polarity Classification** (*pc*): [Pang and Lee 2005], [Go et al. 2009], [Maas et al. 2011], [McAuley and Leskovec 2013], [Rosenthal et al. 2014], [Zhang et al. 2015], [Kaggle 2020d], [Bastan et al. 2020], [Sheng and Uthus 2020]; **Stance Detection** (*sd*): [Kiesel et al. 2019], [Kiesel et al. 2019], [Kawintiranon and Singh 2021], [Kawintiranon and Singh 2021]; **Utility Analysis** (*ua*): [Grano et al. 2017], [Gräßer et al. 2018], [Kaggle 2020b], [Keung et al. 2020], [Kaggle 2020a].

²https://github.com/marilia-cr-silva/nlp_datasets

References

- [Abu Salem et al. 2019] Abu Salem, F. K., Al Feel, R., Elbassuoni, S., Jaber, M., and Farah, M. (2019). FA-KES: A Fake News Dataset around the Syrian War.
- [Alex et al. 2021] Alex, N., Lifland, E., Tunstall, L., Thakur, A., Maham, P., Riedel, C., Hine, E., Ashurst, C., Sedille, P., Carlier, A., Noetel, M., and Stuhlmüller, A. (2021). RAFT: A Real-World Few-Shot Text Classification Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, pages 1–12.
- [Basile et al. 2019] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.
- [Bastan et al. 2020] Bastan, M., Koupaei, M., Son, Y., Sicoli, R., and Balasubramanian, N. (2020). Author’s Sentiment Prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615.
- [Blohm et al. 2021] Blohm, M., Hanussek, M., and Kintz, M. (2021). Leveraging Automated Machine Learning for Text Classification: Evaluation of AutoML Tools and Comparison with Human Performance. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pages 1131–1136.
- [Bojanowski et al. 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- [Chakravarthi 2020] Chakravarthi, B. R. (2020). HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, pages 41–53.
- [Davidson et al. 2017] 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*, pages 512–515.
- [de Gibert et al. 2018] de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 11–20.
- [Demšar 2006] Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, pages 1–30.
- [Demszky et al. 2020] Demszy, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- [Erickson et al. 2020] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. *arXiv:2003.06505*, pages 1–28.
- [Feurer et al. 2015] Feuer, M., Klein, A., Eggenberger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, page 2755–2763.
- [Founta et al. 2018] Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *11th International Conference on Web and Social Media, 2018*, pages 491–500.

- [Gautam et al. 2020] Gautam, A., Mathur, P., Gosangi, R., Mahata, D., Sawhney, R., and Shah, R. R. (2020). #metooma: multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of International AAAI Conference on Web and Social Media*, pages 209–216.
- [Go et al. 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, pages 1–6.
- [Grano et al. 2017] Grano, G., Di Sorbo, A., Mercaldo, F., Visaggio, C. A., Canfora, G., and Panichella, S. (2017). Software Applications User Reviews.
- [Gräßer et al. 2018] Gräßer, F., Kallumadi, S., Malberg, H., and Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In *Proceedings of the 2018 International Conference on Digital Health*, page 121–125.
- [Grosz and Conde-Cespedes 2020] Grosz, D. and Conde-Cespedes, P. (2020). Automatic Detection of Sexist Statements Commonly Used at the Workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: 2020 Workshops*, page 104–115.
- [Guyon et al. 2016] Guyon, I., Chaabane, I., Escalante, H. J., Escalera, S., Jajetic, D., Lloyd, J. R., Macià, N., Ray, B., Romaszko, L., Sebag, M., Statnikov, A., Treguer, S., and Viegas, E. (2016). A brief review of the chlearn automl challenge: Any-time any-dataset learning without human intervention. In *Proceedings of the Workshop on Automatic Machine Learning*, pages 21–30.
- [Hugging Face 2019] Hugging Face (2019). Tweets Hate Speech Detection. Accessed: 2022-04-05.
- [Jigsaw 2018] Jigsaw (2018). Toxic Comment Classification Challenge. Accessed: 2022-04-06.
- [Kaggle 2020a] Kaggle (2020a). Samsung Internal SSD Reviews. Accessed: 2022-04-06.
- [Kaggle 2020b] Kaggle (2020b). Amazon Musical Instruments Reviews. Accessed: 2022-04-06.
- [Kaggle 2020c] Kaggle (2020c). Terrorism And Jihadism Speech Detection. Accessed: 2022-04-06.
- [Kaggle 2020d] Kaggle (2020d). Apple Twitter Sentiment Texts. Accessed: 2022-04-06.
- [Kawintiranon and Singh 2021] Kawintiranon, K. and Singh, L. (2021). Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735.
- [Keung et al. 2020] Keung, P., Lu, Y., Szarvas, G., and Smith, N. A. (2020). The Multilingual Amazon Reviews Corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4563–4568.
- [Kiesel et al. 2019] Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., and Potthast, M. (2019). SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.
- [Maas et al. 2011] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.
- [Mathew et al. 2021] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14867–14875.
- [McAuley and Leskovec 2013] McAuley, J. and Leskovec, J. (2013). Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, page 165–172.
- [Mollas et al. 2020] Mollas, I., Chrysopoulou, Z., Karlos, S., and Tsoumakas, G. (2020). ETHOS: an Online Hate Speech Detection Dataset. *arXiv: 2006.08328*, pages 1–16.

- [Olson and Moore 2016] Olson, R. S. and Moore, J. H. (2016). Tpot: A tree-based pipeline optimization tool for automating machine learning. In *Proceeding of the ICML 2016 AutoML Workshop*, pages 66–74.
- [Pang and Lee 2005] Pang, B. and Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124.
- [Pérez-Rosas et al. 2018] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.
- [Pineau et al. 2020] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Larochelle, H. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). pages 1–22.
- [Reimers and Gurevych 2019] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.
- [Rosenthal et al. 2014] Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment Analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 73–80.
- [Saravia et al. 2018] Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.
- [Shahi and Nandini 2020] Shahi, G. K. and Nandini, D. (2020). FakeCovid - A Multilingual Cross-domain Fact Check News Dataset for COVID-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, pages 1–9.
- [Sheng and Uthus 2020] Sheng, E. and Uthus, D. (2020). Investigating Societal Biases in a Poetry Composition System. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106.
- [Škrlj et al. 2021] Škrlj, B., Martinc, M., Lavrač, N., and Pollak, S. (2021). autoBOT: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, 110(5):989–1028.
- [Sosea et al. 2022] Sosea, T., Pham, C., Tekle, A., Caragea, C., and Li, J. J. (2022). Emotion analysis and detection during COVID-19. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6938–6947.
- [Strapparava and Mihalcea 2007] Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 70–74.
- [Thorne et al. 2018] Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.
- [Thorne et al. 2019] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2019). The FEVER2.0 Shared Task. In *Proceedings of the Second Workshop on Fact Extraction and VERification*, pages 1–6.
- [Torabi Asr and Taboada 2018] Torabi Asr, F. and Taboada, M. (2018). The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity. In *Proceedings of the First Workshop on Fact Extraction and VERification*, pages 10–15.

- [Wang 2017] Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 422–426.
- [Waseem and Hovy 2016] Waseem, Z. and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.
- [Weinzierl and Harabagiu 2022] Weinzierl, M. and Harabagiu, S. (2022). VaccineLies: A Natural Language Resource for Learning to Recognize Misinformation about the COVID-19 and HPV Vaccines. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6967–6975.
- [Zampieri et al. 2019] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420.
- [Zhang et al. 2015] Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, pages 1–9.

Proposta e Avaliação Linguística de Técnicas de Aumento de Dados

Arthur Scalercio¹, Cláudia Freitas²

¹Instituto de Computação – Universidade Federal Fluminense (UFF)

²Departamento de Letras – Pontifícia Universidade Católica (PUC-RJ)

arthurscalercio@id.uff.br, claudiafreitas@puc-rio.br

Abstract. *In Natural Language Processing (NLP), data augmentation consists of creating artificial training data for machine learning through textual transformations, aiming to improve the model’s generalization capabilities and its performance in a range of downstream NLP tasks. Most studies on data augmentation techniques are evaluated based on the performance of the trained model with the artificially generated texts directly in the target task, without concern for linguistically evaluating the quality of the created texts. In this study, we propose two data augmentation techniques, evaluate the linguistic quality of the transformed texts, and demonstrate that the texts are well-constructed linguistically.*

Resumo. *Em Processamento de Linguagem Natural (PLN), aumento de dados consiste na criação artificial de dados para treinamento de modelos de aprendizado de máquina através de transformações nos textos, visando aumentar a capacidade de generalização de modelos e melhorar o desempenho em diversas tarefas de PLN. A maioria dos estudos sobre técnicas de aumento de dados são avaliados pelo desempenho do modelo treinado com os textos artificiais diretamente na tarefa alvo, não se preocupando em avaliar linguisticamente a qualidade dos textos criados. Nesse estudo, propomos duas técnicas de aumento de dados, avaliamos a qualidade linguística dos textos transformados e mostramos que os textos são linguisticamente bem construídos.*

1. Introdução

Nem todo problema de aprendizado é solucionado com aumento de dados (AD). Porém, a quantidade de dados para treinamento tem sido fundamental para o treinamento com sucesso de grandes modelos de linguagem [Radford et al. 2019, Devlin et al. 2019, Brown et al. 2020, Souza et al. 2020]. Há diversas situações em que se mostra vantajoso o uso de dados artificialmente construídos. Por exemplo, [Krishna et al. 2020] e [Lai et al. 2021] mostram que a adoção de paráfrases artificialmente contruídas melhora o desempenho de modelos neurais em diversas tarefas de transferência de estilo, como imitação autoral [Xu et al. 2012] e transferência de sentimento [Li et al. 2018]. A primeira consiste em converter o estilo de uma sentença para o estilo de determinado autor. A segunda consiste em converter a sentença para um sentimento diferente, mas mantendo seu assunto principal.

Normalmente, o sucesso da estratégia de aumento é medido avaliando-se o desempenho do modelo treinado nos dados aumentados em comparação com o modelo treinado nos dados não-aumentados, sem se preocupar se os textos gerados são bem formados. Diferentemente, esse estudo realiza uma avaliação intrínseca dos dados gerados, não avaliando se o aumento beneficia o desempenho de modelos em outras tarefas de PLN.

De forma exploratória, propõem-se, implementam-se e avaliam-se duas transformações para AD. Busca-se manter, linguisticamente, a correção e a naturalidade das frases criadas. Usando a gramática *Universal Dependency* (UD) [de Marneffe et al. 2021], movem-se sintagmas adverbiais que aparecem antes do respectivo verbo que modificam para depois dos blocos de sujeito, verbo e objeto. Usando o mesmo raciocínio, deslocam-se orações subordinadas adverbiais que aparecem antes do respectivo verbo que modificam para o final da frase, após sujeito, verbo e objetos. Três livros da coleção Literatura para Todos (CLPT) foram usados para avaliar os métodos. Excetuando quando o anotador nos fornece uma árvore sintática errada, as frases geradas mostraram-se bem construídas.

2. Método

O advérbio é fundamentalmente um modificador do verbo, podendo também reforçar o sentido de um adjetivo ou de um outro advérbio. Além disso, salienta-se que alguns advérbios aparecem, não raro, modificando toda a oração [Veloso 2013, Costa 2008, Bechara 2001]. Neste último emprego, vêm geralmente destacados no início da oração, de cujos termos se separam por uma pausa nítida, marcada na escrita por vírgula. Quando movidos de lugar, geralmente podem ocupar qualquer outra posição na sentença, e até mesmo separar sujeito-verbo e verbo-objeto. Por outro lado, modificadores adverbiais internos à sentença geralmente não são marcados por pontuação e possuem uma posição mais fixa em relação ao termo que modificam.

Nossa estratégia de AD consiste em mudar de lugar estruturas adverbiais modificadoras cuja posição na sentença seja flexível. Para tanto, duas estruturas com função de advérbio foram escolhidas. Selecionaram-se as frases em que há uma oração subordinada adverbial aparecendo antes do respectivo verbo modificado por ela e as frases em que há um sintagma adverbial aparecendo antes do respectivo verbo modificado por ele. Em ambas as transformações, gera-se uma paráfrase usando as mesmas palavras da sentença de origem, mas movendo a estrutura linguística alvo para depois de sujeito, verbo e objetos.

Para alcançar esse objetivo, anotamos os textos conforme a gramática UD, usando o anotador sintático UDPipe [Straka et al. 2016], que fornece árvores de dependência que nos permitem encontrar as duas estruturas. Os rótulos sintáticos OBL e ADVCL da UD nos trazem, respectivamente, núcleos de sintagmas adverbiais e de orações subordinadas adverbiais. Uma vez encontrados, move-se a palavra correspondente ao núcleo da estrutura juntamente com todos os seus filhos na árvore de dependência para frente da frase, após os termos que constituem o sujeito, o verbo e os objetos. Mais detalhes do algoritmo que faz a translação encontram-se no Apêndice A.1. A árvore de dependência da sentença criada é igual à da sentença de entrada, ou seja, o método proposto não muda nem as palavras existentes na sentença, nem a função sintática exercida por cada uma delas. A ideia da translação do sintagma adverbial está ilustrada na Figura 1. Por último, um pós-processamento dos textos, descrito no Apêndice A.2, é feito para ajustar pontuações.

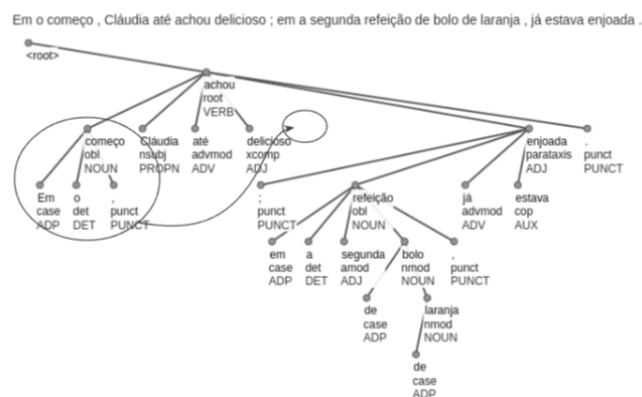


Figura 1. Translação do sintagma adverbial

3. Experimentos e Resultados

Para avaliar nosso método de AD, selecionaram-se três livros da coleção Literatura para Todos (CLPT): *Léo, o pardo*; *Madalena*; e *Tubarão com faca nas costas*. A escolha foi motivada pelo fato de eles serem obras literárias publicamente disponíveis¹ e destinados teoricamente a pessoas com habilidades de leitura iniciais. As estatísticas do conjunto de dados e a quantidade de análises qualitativas realizadas constam na Tabela 1.

Tabela 1. Estatísticas do conjunto de dados e dos experimentos

Rótulo	Livro	#Frases	#tokens	Anotações Avaliadas	Frases Avaliadas (1)	Frases Transformadas	Transformadas avaliadas (2)	Interseção de (1) e (2)
ADVCL	Leopardo	997	25027	20	20	88	20	18
	Madalena	1413	19655	20	20	92	20	20
	Tubarão	650	11678	20	20	99	20	20
	Total	3060	56360	60	60	279	60	58
OBL	Leopardo	997	25027	20	16	220	20	12
	Madalena	1413	19655	20	13	125	20	11
	Tubarão	650	11678	20	15	121	20	13
	Total	3060	56360	60	44	466	60	36

3.1. Análise de erros da anotação automática

Para cada uma das estruturas linguísticas procuradas, calculou-se a precisão do anotador UDPipe. Este anotador morfossintático foi escolhido pela facilidade de uso. Para ambas as estruturas linguísticas analisadas, 60 amostras foram aleatoriamente escolhidas, sendo 20 de cada livro. Para o rótulo ADVCL a precisão foi de 90%, enquanto para o OBL foi de 83.3%. Considerou-se que o anotador acerta se, ao mesmo tempo, classifica corretamente o rótulo sintático UD da palavra (ADVCL ou OBL, conforme o caso) e escolhe corretamente o pai da palavra na árvore sintática. Tabela 2 mostra os erros cometidos pelo anotador conforme essas dimensões. Mais detalhes sobre essa análise no Apêndice A.4.

Tabela 2. Análise da anotação sintática automática

Livro	Orações adverbiais antes do verbo			Sintagmas adverbiais antes do verbo		
	Erro Classificação	Erro Referência	Erro Ref. e Classif.	Erro Classificação	Erro Referência	Erro Ref. e Classif.
Leo-o-pardo	1	1	0	2	0	0
Madalena	1	0	0	1	0	3
Tubarão	1	0	2	2	1	1
Total	3	1	2	5	1	4

¹ <http://www.dominiopublico.gov.br/>

3.2. Análise da qualidade dos dados aumentados

Avaliaram-se qualitativamente 60 frases para cada estratégia de AD, sendo 20 de cada livro. Apesar de a maioria das frases analisadas terem como origem as mesmas cujas anotações foram avaliadas na Subseção 3.1, alguns textos aumentados analisados vieram de frases que não tiveram a anotação avaliada anteriormente, devido a alguns filtros inseridos na seleção das sentenças a serem aumentadas. Detalhes do processo de seleção das frases analisadas encontram-se no Apêndice A.3. Os resultados constam na Tabela 3.

Tabela 3. Análise dos dados aumentados

Livro	Orações adverbiais antes do verbo		Sintagmas adverbiais antes do verbo	
	#Frases Mal-formadas	#Análises	#Frases Mal-formadas	#Análises
Leo-o-pardo	2	20	4	20
Madalena	1	20	6	20
Tubarão	3	20	5	20
Total	6	60	15	60

Dentre as frases analisadas, todas as mal-formadas foram causadas por erros na anotação sintática da frase original. Uma análise das frases mal-formadas encontra-se no Apêndice A.5. Amostras das sentenças bem formadas e analisadas constam no Apêndice A.6. O código que permite a reproducibilidade integral dos experimentos encontra-se disponível².

4. Trabalhos Relacionados

[Bayer et al. 2022] realizou uma pesquisa sobre métodos de AD voltados para a tarefa de classificação e mostrou técnicas que atuam no nível do caractere, da palavra, da frase e do documento, e podem ser baseadas em regras ou em algoritmos de aprendizado de máquina. Dentre as baseadas em regras que criam dados com transformações sintáticas, [Min et al. 2020] melhora o resultado de modelos na tarefa de inferência textual. Para o português brasileiro, [Menezes et al. 2023] mostra que modelos para simplificação textual treinados com dados aumentados apresentam maior generalização quando aplicados a outros domínios. [Şahin and Steedman 2018] cria dados, através da seleção de pedaços de frases e da inversão da ordem de aparecimento de alguns termos sintáticos nas frases. Esses dados artificiais beneficiaram o desempenho de um modelo neural na tarefa de anotação de desambiguação morfossintática (em inglês, part-of-speech (POS) tagging), em diversas línguas com poucos recursos. De forma parecida, nossa abordagem cria dados através de transformações. No entanto, estas são mais complexas e levam em conta as peculiaridades do posicionamento de modificadores adverbiais na língua portuguesa.

5. Conclusão e Trabalhos Futuros

O presente estudo propôs e avaliou linguisticamente duas estratégias para a geração de paráfrases. Embora a amostra analisada seja pequena, a ideia se mostrou promissora e é possível estendê-la para mover outras estruturas linguísticas que possam estar fora da ordem canônica do português, como sujeito e objeto. Vimos que o anotador UDPipe, para sentenças com poucos verbos e sem discurso direto, possui uma boa qualidade. Planeja-se confirmar a efetividade do método avaliando se o aumento beneficia modelos de aprendizado em outras tarefas, como na de simplificação textual, onde frases na ordem canônica e fora da ordem canônica podem ser usadas em um treinamento supervisionado.

²<https://github.com/scalercio/aumento-dados-ud>

Referências

- Bayer, M., Kaufhold, M.-A., and Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.
- Bechara, E. (2001). *Moderna Gramática Portuguesa*. Lucerna.
- 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.
- Costa, J. (2008). *O Advérbio em Português Europeu*. Colibri, Lisboa.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- 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.
- Krishna, K., Wieting, J., and Iyyer, M. (2020). Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.
- Lai, H., Toral, A., and Nissim, M. (2021). Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.
- Menezes, L., Paes, A., and Finatto, M. (2023). Abordagem baseada em aumento de dados para avaliação automática de leitura de dados. *Domínios de Linguagem*, 17:e1721.
- Min, J., McCoy, R. T., Das, D., Pitler, E., and Linzen, T. (2020). Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.
- Şahin, G. G. and Steedman, M. (2018). Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

- 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.
- 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’16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).
- Veloso, R. (2013). *Gramática do Português*, volume 2, chapter Advérbio e Sintagma Adverbial, pages 1569–1684. Fundação Calouste Gulbenkian.
- Xu, W., Ritter, A., Dolan, B., Grishman, R., and Cherry, C. (2012). Paraphrasing for style. In Kay, M. and Boitet, C., editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2899–2914. Indian Institute of Technology Bombay.

A. Apêndice

A.1. Detalhes do Algoritmo de Translação de Estruturas Linguísticas

Para encontrar as frases passíveis de serem transformadas pela nossa técnica, busca-se a palavra que esteja rotulada em UD como OBL (núcleo do sintagma adverbial) ou ADVCL (núcleo da oração subordinada adverbial) e cujo pai dela na árvore de dependência sintática esteja localizado a frente dela. Caso a frase não tenha essa estrutura, nada é feito com ela. Caso alguma dessas estruturas esteja presente, a técnica de aumento será aplicada nela. O núcleo da estrutura juntamente com seus filhos na árvore serão levados para uma posição após o verbo e objetos, caso hajam. Para o sucesso linguístico da sentença gerada, é necessário que a estrutura movida seja inserida não só após as palavras que compõem sujeito, verbo e objetos, mas também após alguns rótulos que indicam palavras que estão vinculadas diretamente ao verbo, tais como verbos de ligação, verbos auxiliares, conjunções, pronomes reflexivos e pontuações. Para englobar tais estruturas, seguem os rótulos em UD, vinculados ao pai da estrutura a ser movida, que devem ser considerados: nsubj, advmod, iobj, obj, obl, advcl, aux, punct, cop, mark, conj, csubj, xcomp, expl. Assim, a estrutura movida é inserida após a ocorrência do último de qualquer desses rótulos.

Para preservar a qualidade dos textos gerados, o algoritmo também não permite que os advérbios modificadores sejam movidos de posições anteriores para posteriores em relação a algumas pontuações (".", ":", "!", "?"). Além disso, durante o processo de translação dos blocos de sintagmas adverbiais e de orações adverbiais, caso haja outro bloco com a mesma função sintática do bloco movido e que tenha índice maior que seu pai na árvore, o bloco movido será inserido antes desse bloco.

Pontua-se que quando o termo a ser movido inicia a frase original, insere-se na sentença criada uma vírgula imediatamente antes da posição onde o termo é inserido. Por fim, quando se busca os sintagmas adverbiais para transformação, não se consideram as palavras anotadas como OBL e que estejam anotadas como pronome relativo (ex: "que"). Trata-se de ajustes que tiveram que ser feitos no código para permitir uma geração de textos adequada. Por não ser alvo do presente estudo, o algoritmo não se preocupou em converter letras maiúsculas para minúsculas e vice-versa.

A.2. Pós-processamento

Após o algoritmo principal, um simples pós-processamento das frases geradas é realizado para ajustar pontuações que passaram a aparecer em sequência. Seguem os ajustes:

- DE , , → PARA ,
- DE , . → PARA .
- DE , ; → PARA ;
- DE , ? → PARA ?
- DE , ! → PARA !

A.3. Detalhes do processo de seleção das frases analisadas

Para a realização da análise de erros do anotador UDPipe, escolheram-se aleatoriamente sentenças dos três livros da CLPT, e nessas sentenças buscaram-se as etiquetas OBL ou ADVCL (de acordo com a estrutura linguística analisada) que apareciam antes do termo

que modificavam. Pontua-se que havia frases que possuíam mais de uma etiqueta OBL na situação procurada. Essa é a razão que fez com que essa análise utilizasse menos de 20 (vinte) frases por livro, apesar de terem sido analisadas 20 estruturas linguísticas por livro (ver Tabela 1).

Para a escolha das frases aumentadas a serem avaliadas linguisticamente, pegaram-se as sentenças geradas a partir das frases que tiveram suas anotações avaliadas. Porém, como o processo de avaliação da anotação usou menos de 20 frases por livro, escolheram-se aleatoriamente as restantes até completar 20 (vinte) frases por livro.

A.4. Detalhes dos erros do anotador automático

Os erros de classificação e de referência constantes na Tabela 2 encontram-se mais detalhados nas Tabelas 4 e Tabelas 5. Percebe-se uma dificuldade do anotador em lidar com discurso direto, o qual deveria ser anotado como PARATAXIS, pois não há uma conexão formal, explícita, entre as duas orações.

Tabela 4. Erros do Anotador - ADVCL

Tipo de Erro	Rótulo Correto	#Qtd	%
Classificação	PARATAXIS	1	16.6%
	ADVMOD	1	16.6%
	XCOMP	1	16.6%
Classif. e	ACL:REL	1	16.6%
Referência	PARATAXIS	1	16.6%
Referência		1	16.6%
Total		6	100%

Tabela 5. Erros do Anotador - OBL

Tipo de Erro	Rótulo Correto	#Qtd	%
Classificação	I OBJ	3	30%
	PARATAXIS	1	10%
	MARK	1	10%
Classif. e	ROOT	3	30%
Referência	PARATAXIS	1	10%
Referência		1	10%
Total		10	100%

A.5. Análise das frases aumentadas mal-formadas

Tabela 6, Tabela 7 e Tabela 8 listam algumas frases mal-formadas usando nosso método para os livros “*Léo, o pardo*”, “*Madalena*” e “*Tubarão com faca nas costas*”, respectivamente. Da análise, nota-se, de um modo geral, que estas sentenças aumentadas são provenientes de frases com discurso direto ou com diversas orações.

A.6. Amostras de frases originais e aumentadas bem-formadas

Tabela 9 e Tabela 10 listam algumas frases aumentadas do livro “*Léo, o pardo*” e consideradas bem-formadas. Tabela 11 e Tabela 12 listam algumas frases bem-formadas do livro “*Madalena*”. E Tabela 13 e Tabela 14 listam algumas frases bem-formadas do livro “*Tubarão com faca nas costas*”.

Tabela 6. Frases mal-formadas do livro “Léo, o pardo”

TIPO	Translação Oração adverbial
Original	“ Mãe , a professora disse que toda mãe deixa de comer pra dar pro filho , se for preciso , igual alguns bichos ” .
Aumentada	“ Mãe , a professora disse que toda mãe deixa de comer pra dar pro filho , igual alguns bichos se for preciso ” .
Original	“ Como , se eu não fiz colégio bom e não sei química e nem física ? ” Ele jogou a deixa , primeiro igual o professor Lázaro .
Aumentada	“ Como , jogou se eu não fiz colégio bom e não sei química e nem ? ” Ele física a deixa , primeiro igual o professor Lázaro .
TIPO	Translação Sintagma adverbial
Original	Pra gente ela sempre aconselhava , “ não chora tanto , menino , que lágrima é uma coisa que seca , quando a alma mais precisar de ela , aí não desce mais ” .
Aumentada	ela sempre aconselhava , Pra gente , “ não chora tanto , menino , que lágrima é uma coisa que seca , quando a alma mais precisar de ela , aí não desce mais ” .
Original	Em a primeira calça ela foi meio sem jeito , mediu a parte de a frente , fez a parte de trás maior por causa de a bunda , emendou as duas , o cliente gostou e ela passou a fazer roupa pra fora .
Aumentada	ela foi meio , Em a primeira calça sem jeito , mediu a parte de a frente , fez a parte de trás maior por causa de a bunda , emendou as duas , o cliente gostou e ela passou a fazer roupa pra fora .

Tabela 7. Frases mal-formadas do livro “Madalena”

TIPO	Translação Oração adverbial
Original	Porque pensar ela podia , e pensando ela podia tudo .
Aumentada	podia , Porque pensar ela , e pensando ela podia tudo .
TIPO	Translação Sintagma adverbial
Original	- Mais ou menos , mainha - e ele olhou para Dulce .
Aumentada	- mainha Mais ou menos , - e ele olhou para Dulce .
Original	- É muito linda , Dinda ... - e , com a melhor cara de cachorro de porta de açougue que sabia fazer - Um dia me dá ela ?
Aumentada	- É muito linda , Dinda ... - e - Um dia me dá ela , com a melhor cara de cachorro de porta de açougue que sabia fazer ?
Original	Quando amanheceu , viu Francisco dormindo em o banco de trás de o carro de dona Madalena , o mesmo banco onde tantas e tantas vezes ... Bateu em o vidro .
Aumentada	Quando amanheceu , viu Francisco dormindo em o banco de trás de o carro de dona Madalena , o mesmo banco onde Bateu tantas e tantas ... vezes em o vidro .
Original	E nem por ela ele fez alguma coisa , quando Nelson a entregou a Manuel em troca de uma braçada de terras .
Aumentada	E ele fez alguma coisa , quando Nelson a entregou a Manuel em troca de uma braçada de terras nem por ela .

Tabela 8. Frases mal-formadas do livro “Tubarão com faca nas costas”

TIPO	Translação Oração adverbial
Original	E assim fui , feliz , por um pedaço de o percurso , com aquela criança pequena me confortando mais de o que eu a ela , até que - surpresa - a mãe decidiu descer bem antes de onde disse que desceria .
Aumentada	E assim fui , feliz , por um pedaço de o percurso , com aquela criança pequena me confortando mais de o que eu a ela , até que a mãe decidiu descer bem antes de onde disse que desceria - surpresa - .
Original	Só deu tempo de ele parar o ônibus . “Não cheguei a conhecer os pais de Santiago , mas toda vez que lhe faço uma visita , seja para jogarmos xadrez ou para discutirmos um novo texto , ele dá um jeito de falar em ambos , principalmente em a mãe , que era com quem ele mais se identificava .
Aumentada	Só deu tempo de ele parar o ônibus . “Não cheguei a conhecer os pais de Santiago , ele dá um jeito de falar em ambos , principalmente em a mãe , que era com quem ele mais se identificava , mas toda vez que lhe faço uma visita , seja para jogarmos xadrez ou para discutirmos um novo texto .
TIPO	Translação Sintagma adverbial
Original	É graças a esse bichano que tenho para mim que nenhum animal de estimação é demasiadamente esperto , senão não aceitariam viver em uma família como a nossa .
Aumentada	É que tenho graças a esse bichano para mim que nenhum animal de estimação é demasiadamente esperto , senão não aceitariam viver em uma família como a nossa .
Original	Em suma : o amor deve estar incrustado em tudo e em todos .
Aumentada	deve , Em : o amor suma estar incrustado em tudo e em todos .

Tabela 9. Frases bem-formadas do livro “Léo, o pardo”

TIPO	Translação oração adverbial
Original	Revoltadas com a condenação , falaram pra juíza : “ então manda matar de uma vez ” .
Aumentada	falaram pra juíza , Revoltadas com a condenação : “ então manda matar de uma vez ” .
Original	”Pra fazer cinema , primeiro se aprende fotografia , literatura , se vê muito filme , música .
Aumentada	”primeiro se aprende fotografia , literatura Pra fazer cinema , se vê muito filme , música .
Original	Voltando de um trabalho como monitor em a biblioteca de o Instituto de Matemática , reencontrei a Sabrina , professora de o curso de cinema de Santo André , e recebi de ela os parabéns por ter passado em a USP .
Aumentada	reencontrei a Sabrina , professora de o curso de cinema de Santo André , Voltando de um trabalho como monitor em a biblioteca de o Instituto de Matemática , e recebi de ela os parabéns por ter passado em a USP .
Original	A máquina de lavar roupas quebrou em a única vez que eu usei ; ”se quebrar de novo será descontado de o seu salário ” .
Aumentada	A máquina de lavar roupas quebrou em a única vez que eu usei ; “ será descontado de o seu salário se quebrar de novo ” .
Original	Sentado em o próximo 109 ônibus , fiquei pensando o que aconteceria se não eu fosse de a USP e ainda estivesse desempregado .
Aumentada	fiquei pensando o que aconteceria se não eu fosse de a USP e ainda estivesse desempregado , Sentado em o próximo 109 ônibus .
Original	Em o trabalho , a Dona Nega gritava de lá : “ alguém quer bala ? ” E a Maria de o Antônio Joana , depois de aceitar , gritava em o outro canto , cuspindo : “ cruz credo , nunca comi bala de espinho ! ” E a Dona Nega caía em a risada : “ é de coco , criatura ” .
Aumentada	Em o trabalho , a Dona Nega gritava de lá : “ alguém quer bala ? ” E a Maria de o Antônio Joana gritava em o outro canto , depois de aceitar , cuspindo : “ cruz credo , nunca comi bala de espinho ! ” E a Dona Nega caía em a risada : “ é de coco , criatura ” .

Tabela 10. Frases bem-formadas do livro “Léo, o pardo”

TIPO	Translação Sintagma adverbial
Original	Anos mais tarde , com esse tal Cambraia já morto , o filho de ele mais moço , o Geirson Cambraia , descobriu um monte de papel amarelado em uma gaveta e quase morreu de desgosto ; eram os documentos de tudo quanto era terra alheia que se passava por terra de os Cambraia .
Aumentada	com esse tal Cambraia já morto , o filho de ele mais moço , o Geirson Cambraia , descobriu um monte de papel amarelado em uma gaveta , Anos mais tarde , e quase morreu de desgosto ; eram os documentos de tudo quanto era terra alheia que se passava por terra de os Cambraia .
Original	Enquanto isso , em a nossa curiosidade que nunca morria , ela ia contando detalhes de o tempo antigo .
Aumentada	em a nossa curiosidade que nunca morria , ela ia contando detalhes de o tempo antigo , Enquanto isso.
Original	Em a USP , a maioria de os homens e mulheres negras , e são muitos , vestem roupas vermelhas , verdes , alaranjadas , azuis ou pretas e cortam gramas , capinam jardins , guardam portas e servem comidas em bandejas pros alunos brancos , em o restaurante central , ou laboram em eternas construções e depois vão pra casa , ali de o lado .
Aumentada	a maioria de os homens e mulheres negras , e são muitos , vestem roupas vermelhas , verdes , alaranjadas , azuis ou pretas , Em a USP , e cortam gramas , capinam jardins , guardam portas e servem comidas em bandejas pros alunos brancos , em o restaurante central , ou laboram em eternas construções e depois vão pra casa , ali de o lado .
Original	Com o cargo ganharia mil e seiscentos reais e faria faculdade de cinema .
Aumentada	ganharia mil e seiscentos reais , Com o cargo e faria faculdade de cinema .

Tabela 11. Frases bem-formadas do livro “Madalena”

TIPO	Translação oração adverbial
Original	Enquanto Madalena se vestia , Rubina puxou sua maleta de couro de baixo de a cama e foi enfiando em ela o conteúdo de a trouxa : uma meia com algum dinheiro , dois vestidos , uma camisola , um casaco , toalhinhas , cueiros , dois conjuntos de Chiquinho .
Aumentada	Rubina puxou sua maleta de couro de baixo de a cama , Enquanto Madalena se vestia , e foi enfiando em ela o conteúdo de a trouxa : uma meia com algum dinheiro , dois vestidos , uma camisola , um casaco , toalhinhas , cueiros , dois conjuntos de Chiquinho .
Original	E , apesar de nunca ter destratado Francisco , Maria tampouco gostava de esse namoro ; sempre que podia , trazia a a tona esse assunto : ”Cuidado com esse Francisco .
Aumentada	E Maria tampouco gostava de esse namoro , apesar de nunca ter destratado Francisco ; sempre que podia , trazia a a tona esse assunto : “ Cuidado com esse Francisco .
Original	Além de ser a encarregada de todos os não disciplinadores , Dulce era , a os olhos de Cláudia , infinitamente menos interessante que Madalena .
Aumentada	Dulce era , a os olhos de Cláudia , infinitamente menos interessante que Madalena , Além de ser a encarregada de todos os não disciplinadores .
Original	Se estava se separando por a segunda vez era por uma fatalidade .
Aumentada	era por uma fatalidade , Se estava se separando por a segunda vez .
Original	Não perdia tempo discutindo , mas tinha lá suas idéias , e quando cismava com uma coisa , não sossegava até conseguir o que queria .
Aumentada	Não perdia tempo discutindo , mas tinha lá suas idéias , e não sossegava quando cismava com uma coisa , até conseguir o que queria .

Tabela 12. Frases bem-formadas do livro “Madalena”

TIPO	Translação Sintagma adverbial
Original	Em a época , Dulce ainda pesava menos de quarenta quilos por a depressão que havia sofrido com a morte repentina de o pai , ocorrida dois anos antes de o namoro começar .
Aumentada	Dulce ainda pesava menos de quarenta quilos , Em a época , por a depressão que havia sofrido com a morte repentina de o pai , ocorrida dois anos antes de o namoro começar .
Original	Em o final de cada mês , eles deixavam todo o salário em a venda de a própria fazenda , e em o dia seguinte a o pagamento , fatalmente , já começavam outra dívida .
Aumentada	eles deixavam todo o salário , Em o final de cada mês , em a venda de a própria fazenda , e em o dia seguinte a o pagamento , fatalmente , já começavam outra dívida .
Original	Aquela noite em que Francisco foi arrancado de a cama que dividia com a mãe por um sujeito imenso - ou , por o menos , imenso para os quatro anos que ele tinha - que caiu em cima de ela .
Aumentada	Aquela noite em que Francisco foi arrancado de a cama que dividia com a mãe por um sujeito imenso - ou imenso , por o menos , para os quatro anos que ele tinha - que caiu em cima de ela .
Original	Porque , em casa , a avó não dava trela para ninguém - saía , trabalhava , voltava , viajava , saía de novo .
Aumentada	Porque a avó não dava trela , em casa , para ninguém - saía , trabalhava , voltava , viajava , saía de novo .
Original	- Em a horta - respondeu , recolhendo a garrafa de cachaça e os copos ; verificou se o vidro de o tampo de a mesinha não estava trincado .
Aumentada	respondeu , recolhendo a garrafa de cachaça e os copos ; - Em a horta - verificou se o vidro de o tampo de a mesinha não estava trincado .

Tabela 13. Frases bem-formadas do livro “Tubarão com faca nas costas”

TIPO	Translação oração adverbial
Original	Quando o ônibus chegou , ela pediu (ainda ignorando Maira e demais pessoas) que a ajudasse com o pequeno .
Aumentada	ela pediu , Quando o ônibus chegou , (ainda ignorando Maira e demais pessoas) que a ajudasse com o pequeno .
Original	Antes de perdê-lo de vista , vejo o inteiramente só , em o meio de a rua .
Aumentada	vejo o inteiramente só , em o meio de a rua , Antes de perdê-lo de vista .
Original	Passadas duas ou três cervejas e umas boas risadas , começamos a imaginar o que teria acontecido a a moça , dona de o decote .
Aumentada	começamos a imaginar o que teria acontecido a a moça , dona de o decote , Passadas duas ou três cervejas e umas boas risadas .
Original	Quem o encontrou foi meu pai , que , após comerem , tinha saído para caminhar e conhecer a propriedade que o amigo comprara a fim de dar um pouco de paz a a família .
Aumentada	Quem o encontrou foi meu pai , que tinha saído , após comerem , para caminhar e conhecer a propriedade que o amigo comprara a fim de dar um pouco de paz a a família .
Original	Aproveitando se de esse descuido , um pequeno tumor começou a trabalhar em o seu ardiloso intento .
Aumentada	um pequeno tumor começou a trabalhar em o seu ardiloso intento , Aproveitando se de esse descuido .

Tabela 14. Frases bem-formadas do livro “Tubarão com faca nas costas”

TIPO	Translação Sintagma adverbial
Original	Para tu teres uma idéia , em a mesma época em que o Mariozinho morreu , o Élcio , um grande camarada que tu não chegaste a conhecer , foi morar em Florianópolis com a mulher .
Aumentada	Para tu teres uma idéia , o Élcio , um grande camarada que tu não chegaste a conhecer , foi morar em a mesma época em que o Mariozinho morreu , em Florianópolis com a mulher .
Original	Para sorte não de o gato , morávamos em uma casa de dois andares .
Aumentada	morávamos , Para sorte não de o gato , em uma casa de dois andares .
Original	E , com uma imagem clara de a minha mãe em a saudade , contei uma história de como (quase) me perdi .
Aumentada	E contei uma história de como (quase) me perdi , com uma imagem clara de a minha mãe em a saudade .
Original	Em seguida , tomo um café , volto para o quarto e leio o que me der vontade .
Aumentada	tomo um café , Em seguida , volto para o quarto e leio o que me der vontade .
Original	Dia de esses , quando fomos a Rio Grande , cidade onde nasceste , enquanto me esperavas engraxar os sapatos para a formatura a a qual havíamos sido convidados , o engraxate , um senhor que regulava em idade contigo , em os colocou a par de o que acontecera a um semnúmero de pessoas , todas conhecidas tuas .
Aumentada	quando fomos a Rio Grande , cidade onde nasceste , enquanto me esperavas engraxar os sapatos para a formatura a a qual havíamos sido convidados , o engraxate , um senhor que regulava em idade contigo , em os colocou , Dia de esses , a par de o que acontecera a um semnúmero de pessoas , todas conhecidas tuas .

Semantic Textual Similarity for Abridging Clinical Notes in Brazilian Electronic Health Records

Lucas T. Bandeira¹, Bernardo S. Consoli¹, Renata Vieira², Rafael H. Bordini¹

¹School of Technology – Pontifical Catholic University of Rio Grande do Sul
Porto Alegre – RS – Brazil.

²School of Technology – University of Évora
Évora – Portugal.

l.treviso@edu.pucrs.br, bernardo.consoli@acad.pucrs.br,
renata.v@uevora.pt, rafael.bordini@pucrs.br

Abstract. *With the growing importance of the use of information from electronic patient records in the development of machine learning models, there is also a need for a holistic understanding of those records, in particular abridging the clinical notes so that important information is used in the training process without the repetition that is commonly found in such notes. This paper presents the pre-processing of clinical notes from the BRATECA Dataset, a Brazilian tertiary care data collection, aiming at removing repeated information resulting from the interaction between healthcare providers and patients, considering assigned values of semantic similarity between sentences in clinical notes.*

1. Introduction

In the Artificial Intelligence field, there is significant interest of researchers in developing systems capable of supporting decision making in the healthcare domain [Shamout F 2021]. However, some of the data categories included in the electronic patient records have several characteristics that make their use difficult, which highlights the need to maintain better databases. Although, the most used databases for these purposes are formed by data extracted from hospital records in English-speaking countries, which does not represent the clinical reality in Brazil.

The BRATECA Dataset is one of the few national resources for the development of research projects in computational medicine. In this study, we contribute in that direction by abridging the clinical notes, in particular to remove repeated information, since it is customary for healthcare providers in Brazil to repeat known conditions and ongoing treatments when writing clinical notes. Nevertheless, this repetition obstructs the training of accurate machine learning models for clinical tasks, introducing biased inaccuracies rooted in duplicated content. Our paper outlines the experiments conducted to identify and eliminate potentially repetitive sentences from the clinical notes within BRATECA.

The remainder of this work is organized as follows: Section 2 describes previous work on BRATECA, semantic similarity and word embeddings; Section 3 describes the neural network used; Section 4 describes the data resources used; Section 5 describes the process of abridging clinical notes; Section 6 provides final considerations on the research developed.

2. Related Work

The method identified to generate meaningful data for an AI model considering the heterogeneity of information is through word embeddings, which are vector representations of words in a multidimensional space. As the semantic value of a word is also mapped, it can be inferred that it is possible to perform calculations to discover how similar two sentences are. Thus, studies involving the application of word embeddings were considered, in addition to databases with clinical and general domain resources in Portuguese.

[Consoli et al. 2022] proposed a new collection of Brazilian clinical data containing more than 70,000 admissions, representing a total of more than 2.5 million clinical notes in free text, aiming to create a dataset with Brazilian clinical information.

[Schneider et al. 2020] presents a BERT model trained on clinical texts from electronic medical records of Brazilian hospitals and texts from the biomedical literature. BioBERTpt is the result of transferring knowledge encoded in a multilingual BERT model to a corpus of clinical and biomedical data.

The research conducted by [Mutinda et al. 2021] is noteworthy for creating a dataset of Japanese clinical records through an approach that evaluates the semantic similarity between clinical notes using BERT. The raw text of the documents was first divided into sentences, and a new data collection was created by combining all possible sentence pairs.

Finally, the work carried out by [Real 2021] should be highlighted, aiming to offer a new benchmark for computational semantic tasks in Portuguese, by providing a dataset composed of pairs of sentences annotated with semantic similarity indexes.

3. Neural Network

The developed model consists of fine-tuning BioBERTpt to generate a value for a regression task. In brief, we used the base model in the embedding layer of the selected architecture to create vector representations of the clinical text inputted into the model. To consider both directions of the sequence of values generated by BERT, we added a bidirectional layer during training of the developed model. Next, a Max Pooling layer is applied to filter the numerical vector used as word embeddings. Three Dense Layers are then added, with a Dropout layer between them to randomly turn off nodes and prevent overfitting. The last Dense Layer is constructed with only one node, which generates a single value as the output of the regression task. In this case, the value represents a similarity index between a pair of sentences, ranging from 1 (for sentences containing extremely different information) to 5 (for sentences with practically equal information).

The input text is prepared for the model using bioBERTpt's pre-trained Tokenizer, which employs a WordPiece approach to convert sentences into words and subwords represented by ids. Special tokens like [SEP] are incorporated to signify sentence endings, while the [PAD] token is used to standardize input sizes. The Tokenizer processes two sentences as input, generating a two-dimensional vector containing input ids and attention masks. It's important to highlight that the model creation and data processing script were developed using Python, utilizing libraries including TensorFlow, Keras and spaCy. The model's performance evaluation in the task was executed using Pearson's correlation coefficient, which yielded a score of 0.73.

4. Resources

This work required the use of linguistic resources that cover both the health domain and the general domain. How these resources were used is explained in more details below.

4.1. Work with ASSIN 2

ASSIN 2 is a shared task in the field of natural language processing, which focused on identifying the semantic similarity between pairs of sentences written in Portuguese. Therefore, this dataset was chosen to train the different tested models during the project, as it is one of the few resources in Portuguese with annotated semantic similarity values. As the developed model needed to be trained for a semantic similarity task, we searched for data with similarity values assigned to it. This way, it would be possible to later apply the trained model on clinical data.

4.2. Work with BRATECA

BRATECA presents its information in a free text format, making it highly unstructured for natural language processing tasks. Thus, to eliminate repeated information and compare sentences, it was necessary to process the data by transforming them into sentences that are understandable to a model.

The project's focus is to create patient representations from heterogeneous data, which requires structuring. Due to the free-text format of the records, a pre-trained pipeline in Portuguese from the spaCy library was used to preprocess the records, as well as regular expressions to remove special characters and clean the patient records. After splitting the clinical records into sentences, semantic similarity values were computed by using the previously trained model based on BioBERTpt and ASSIN 2 Dataset. To create patient representations, the clinical records were structured by removing special characters, splitting them into sentences, and eliminating duplicate information. The abridging process was performed by defining the Cartesian product between the sentences and assigning a similarity index to remove highly similar sentences. By processing BRATECA in this way, a new dataset was created that represents a patient's clinical records in a format more suitable for use with AI models.

5. Results

Patients were selected from 4 ranges of clinical records, namely 10, 100, 500, and 1000 records, to gather information on the percentage of repeated information in different hospitalization scenarios. This was done to determine how much information is repeated for patients who are hospitalized for a few days and patients who are hospitalized for a few months.

To create patient representations, all clinical records for a given patient were first selected. The first clinical note was split into sentences and all sentences were added to a new dataset because they contained new information. For subsequent records, the text was also split into sentences and each sentence was compared to the information already in the dataset. If the similarity index between them was greater than or equal to 4, the new sentence was ignored because it duplicated information that had already been obtained.

Thus, in addition to structuring the clinical notes into less heterogeneous sentences, it was possible to eliminate the repeated information present in the clinical records

and consequently significantly reduce the number of clinical notes needed to represent the hospitalization of a patient. Table 1 shows the reduction in the number of clinical notes after eliminating the repeated information using the trained model, varying from 55.45% to 71.45% across the four groups of patients. Furthermore, we checked how many of those sentences were exactly the same for each patient. To achieve this, we compared the Unicode values and calculated the percentage of sentences that were identical. Table 2 presents the results of this comparison, which revealed that a substantial proportion of sentences in clinical notes were duplicates to previous sentences.

Table 1. Clinical notes with similarity less than 4

Patient Id	Clinical Records			
	Clinical Notes	Sentences Before	Sentences After	Reduction
17	10	110	49	55.45%
293	97	661	264	60.06%
76	469	1836	678	63.07%
668	929	4463	1274	71.45%

Table 2. Clinical notes with equal sentences

Patient Id	Clinical Records			
	Clinical Notes	Sentences Before	Sentences After	Equal Sent.
17	10	110	72	34.55%
293	97	661	383	42.06%
76	469	1836	1210	34.10%
668	929	4463	3011	32.54%

5.1. Discussions

The study emphasizes the importance of eliminating redundant information from patient records. The findings suggest that this can significantly decrease the number of clinical notes needed to document a patient’s hospitalization. By analyzing the semantic similarity of each sentence in a patient’s hospitalization record, it was possible to achieve a 70% reduction in the number of sentences while preserving the same informational value as the original record.

It’s important to highlight that the performance of the developed model was hindered in some cases due to grammatical and structural errors in the clinical notes. These errors made it challenging to understand drug names and updates on treatments, particularly when records included abbreviations and lacked white spaces, hindering the model’s ability to extract meaningful information from some clinical notes.

6. Conclusion

This study tested the application of an architecture based on Word Embeddings from BioBERTpt that was fine-tuned for a regression task. The goal was to generate a semantic similarity index between sentences of clinical notes in Portuguese to reduce the heterogeneity of the existing information in the records of the BRATECA. We found that the information presented in the clinical records is often repeated countless times during the patient’s hospital stay, regardless of its length. Thus, it is worth noting that we developed a method to reduce the number of sentences to be processed by an artificial intelligence model while preserving all the information that existed before creating the patient representations. This approach would reduce the computational cost while using BRATECA data.

References

- Consoli, B., dos Santos, H. D. P., Ulbrich, A. H. D. P. S., Vieira, R., and Bordini, R. H. (2022). BRATECA (Brazilian tertiary care dataset): a clinical information dataset for the Portuguese language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5609–5616, Marseille, France. European Language Resources Association.
- Mutinda, F., Yada, S., Wakamiya, S., and Aramaki, E. (2021). Semantic textual similarity in japanese clinical domain texts using bert.
- Real, L., F. E. G. O. H. (2021). The assin 2 shared task: A quick overview. *Methods Inf Med*.
- Schneider, E., Souza, J., Knafou, J., Copara, J., Oliveira, L., Gumiel, Y., Ferro Antunes de Oliveira, L., Teodoro, D., Paraiso, E., and Moro, C. (2020). Biobertpt – a portuguese neural language model for clinical named entity recognition. pages 65–72. Association for Computational Linguistics.
- Shamout F, Zhu T, C. D. (2021). Machine learning for clinical outcome prediction. volume 14, pages 116–126. Institute of Electrical and Electronics Engineers Inc.

Towards analysis on textual inference at ASSIN-2 dataset

Felipe O. da Silva¹, Giovana Meloni Craveiro¹,
Vinícius F. da Silva¹, Vinícius João de Barros Vanzin¹

¹Institute of Mathematics and Computer Sciences
University of São Paulo (USP) – São Carlos, SP – Brazil

{felipe.oliveiras2000}@gmail.com,
{giovana.meloni.craveiro, vfsilva, vinicius.vanzin}@usp.br

Abstract. *In this article, we conduct a preliminary analysis of different methods to address the Textual Entailment Recognition (RTE) task in Portuguese. We use the ASSIN-2 dataset as a benchmark to evaluate our models. Our work combines various textual representation approaches, including bag of words and word embeddings, with machine learning models. Additionally, we present a rule-based approach. Our highest performance was achieved by the BERTimbau-large model fine-tuned on ASSIN-2, which attained an F1 score of 0.89%, positioning it just 1% below the current state-of-the-art. Our ongoing experiment aims to combine our different approaches to leverage their full potential.*

1. Introduction

Textual Entailment Recognition (RTE), also known as Natural Language Inference (NLI), is the NLP task of determining whether one sentence (premise) entails another (hypothesis). Approaches used for NLI include earlier symbolic and statistical methods to more recent deep learning approaches [Bowman and Zhu 2019]. In the last few years, there has been fast progress on the task [Bowman et al. 2015, Chen et al. 2018] with studies of new model architectures aimed at improving performance on benchmarks as well as at increasing the number of large datasets for evaluating these systems [Williams et al. 2018].

There is a scarcity of datasets on Portuguese for NLI. The ASSIN-2 is a curated dataset proposed at the STIL 2019 conference as an effort to set a new computational semantic benchmark for Portuguese. It contains 10,000 samples of annotated data, divided into balanced portions. The training set contains 6500 sentence pairs, the validation set is composed of 500 pairs, and the test set consists of 2448 pairs [Real et al. 2020]. An example from the data set is shown in Table 1.

Table 1. Example of ASSIN-2 data

Premise	Hypothesis	NLI
Alguns animais estão brincando selvagememente na água	Alguns animais estão brincando na água	Entails
Um avião está voando	Um cachorro está latindo	None

This work uses the ASSIN-2 dataset to analyze and compare the performance of diverse classification approaches for NLI in Portuguese. It combines the representation formats of word embeddings and bag of words with machine learning algorithms. It uses Logistic Regression with L1 and L2 regularization, Random Forests,

and eXtreme Gradient Boosting (XGBoost) for the former [Pedregosa et al. 2011] and CatBoost [Prokhorenkova et al. 2018], a Bi-directional recurrent neural network [Schuster and Paliwal 1997] and BERTimbau [Souza et al. 2020], for the latter. Finally, it includes a rule-based symbolic approach.

2. Methodology

2.1. Symbolic approach

The symbolic approach was inspired by the annotation guidelines for ASSIN-2 [Real et al. 2020], which direct towards verifying whether expressions from both sentences could refer to the same extralinguistic elements or whether an expression from one sentence could have a hypernymy relationship with an expression from the other sentence.

Our rule-based method assumes that the sentences must be similar and that if one of them contains a negation, the other should also contain one. Additionally, it assumes that a longer sentence typically carries more specifications than a shorter one and that a more general sentence could be entailed by a more specific one, but not vice versa.

Hence, the classifier declares that sentence A entails sentence B if sentence A is longer than sentence B, their similarity rate is greater than fifty percent, and either no sentence contains the negation term “não” or both sentences do. The similarity rate between them is measured by dividing the number of words that are common in both sentences by the length of sentence B. Preprocessing is based on lowercasing and removing accents.

As this approach does not consider the semantics of different terms and thus cannot identify relationships of synonymy and hypernymy among different words, it cannot be considered a method that appropriately addresses the NLI task. At this stage, it is intended as a baseline for the minimal performance that the other methods should achieve.

2.2. Bag of words

In the approach that uses *Bag of Words* to represent sentences, different techniques were experimented separately and in combination.

The baseline method is the traditional *Bag of Words* [Zhang et al. 2010], which transforms texts into attribute-value tables by calculating the frequency with which words occur in the texts. This technique creates a single matrix for the entire data set, in which each line i contains the number of times that each word j occurred for the i -th sentence in the database. Similarly, the approach based on *Term Frequency-Inverse Document Frequency* (TF-IDF) [Das and Chakraborty 2018], uses attribute-value tables, but also normalizes the frequency of terms in a document, increasing the relevance of rare words. Additionally, an n -grams strategy, which aims to include the context of adjacent words instead of the single referred term, is tested as an alternate configuration.

Aiming to reduce the dimensionality of the training set, we also tested adding *Principal Component Analysis* (PCA) [Shlens 2014] with a varying number of components among 3, 4, 5, 10, 100, and 500. As a result, we had 2305 dimensions for the traditional *BOW* method and 33516 dimensions for approaches using 1 to 3 n -grams.

The cited representation formats and techniques are applied to the data set, pre-processed by removing *stopwords*, and fed into machine learning algorithms based on

Logistic Regression, Random Forests, and eXtreme Gradient Boosting (XGBoost), which are offered by scikit-learn [Pedregosa et al. 2011].

For each combination of hyperparameters, the models are fine-tuned with Grid Search and k-fold cross-validation with $k = 5$ is applied. F1 measure is used to select the best model. It is assumed that no significant weight differences exist between False Positive (FP) and False Negative (FN) errors for this task. It is also the metric used to evaluate RTE systems in the ASSIN-2 benchmark [Real et al. 2020]. A total of 150 different configurations are used, with distinct representation approaches, resulting in 750 predictive models.

2.3. Word embeddings

In this semantic representation format, the NILC pre-trained embeddings [Hartmann et al. 2017] are combined with machine learning classifiers.

1. The first strategy uses the CatBoost algorithm [Prokhorenkova et al. 2018]. Pre-processing consists of normalization to lowercase words and concatenation of premise and hypothesis with a separation token “[SEP]”, without removing stop-words. The sentence is represented by the sum of the individual embedding vectors of each word. Tests are conducted with the embeddings *word2vec* CBow of 100 dimensions and *Glove* Skip-gram of 300 dimensions [Mikolov et al. 2013].
2. The second technique employs a bidirectional recurrent neural network (BRNN) [Schuster and Paliwal 1997]. Each sentence was preprocessed with the techniques described in [Hartmann et al. 2017]. The model was trained with the *Adam* optimization algorithm using at most 25 epochs and a batch of 128 samples. 21 models were trained, varying embeddings (*word2vec skip-gram*, *word2vec CBoW*, *wang2vec skip-gram*, *wang2vec CBoW*, *FastText skip-gram*, *FastText CBoW* e *Glove*) and number of dimensions (50, 300 and 1000).
3. The third method uses BERTimbau [Souza et al. 2020], a Brazilian Portuguese language model, trained on the brWaC corpus [Wagner Filho et al. 2018], fine-tuned [Howard and Ruder 2018] to the RTE task. There are two versions of pre-trained models: one with 12 layers of encoders, 110 million parameters, and 768 dimensions; and one with 24 layers of encoders, 335 million parameters, and 1024 dimensions. The same hyperparameters are used for both versions. The maximum token sequence length is set at 128, the maximum number of epochs is 4, and the batch size is 16 for training and 64 for validation. The remaining hyperparameters were not modified and we use the tokenizer from the pretrained model.

3. Results and Discussion

Among the different strategies used to tackle RTE, several configurations were tested. Table 2 exhibits the results of the ones that obtained the highest scores. For each textual representation method, our code and experiments are openly available at repository [Oliveira da Silva et al. 2023], facilitating replication of results.

Our symbolic approach is designed solely with rules that do not attempt to capture semantic relationships among different words. Despite this aspect, it achieves the remarkably high $F1$ score of 0.71%, given its simplicity. This indicates that it is either

Table 2. Result of the best models

Set	Method	Metrics			
		F1	Precision	Recall	Accuracy
Train	BOW	0.94	0.92	0.95	0.93
	BERTimbau-large	0.96	0.97	0.96	0.96
	Symbolic	0.70	0.75	0.67	0.72
Validation	BOW	0.88	0.87	0.89	0.88
	BERTimbau-large	0.96	0.96	0.96	0.96
	Symbolic	0.72	0.75	0.69	0.73
Test	BOW	0.77	0.68	0.88	0.73
	BERTimbau-large	0.89	0.90	0.89	0.89
	Symbolic	0.69	0.74	0.65	0.71

a promising approach or that the data set used to test the experiments is too simplistic to reflect the complexity of the task in real-world examples. The other approaches are expected to outperform this method.

Our fine-tuned model that uses BERTimbau-large indeed reaches an $F1$ score of 0.89%, only one percentage point away from the current state-of-the-art in the RTE task - represented by a BERTimbau-large trained by Neuralmind [Souza et al. 2020] which achieved 0.90%. However, it is important to perform a qualitative analysis of its misclassifications. Table 3 shows an example from the test set in which our BERTimbau model misclassifies.

Table 3. Example of BERTimbau misclassification

Premise	Hypothesis	NLI
um palhaço está cantando no palco e pessoas estão dançando	uma pessoa fantasiada de palhaço está cantando	Non-Entailment

In the given example, the premise says “um palhaço está cantando no palco” and the hypothesis says “uma pessoa fantasiada de palhaço está cantando”. Although the data set classifies this sentence as non-entailment, “a clown” could be considered as equivalent to “a person dressed as a clown”. Therefore, the model seems to be a solid solution, but it is reasonable to further analyze its misclassifications to ensure its robustness and to understand how to enhance it. Nevertheless, its greatest disadvantage is that its computational cost and complexity are significantly greater than those of the other methods.

Our best combination of a BOW method - without PCA and without TF-IDF - achieves an $F1$ measure of 0.73%, which is markedly lower than that of our best BERTimbau model, but its recall is only 0.01% below our BERTimbau’s recall, which is notable given its significantly lower computational cost compared to the BERTimbau models.

Given the computational cost and accessibility disadvantages of our BERTimbau model and its performance advantage compared to our other methods, our ongoing work aims to refine and combine our methods, resulting in a Neuro-symbolic approach for Portuguese textual inference that considers all linguistic features necessary to properly address NLI, while remaining accessible and competitive with state-of-the-art models.

References

- [Bowman and Zhu 2019] Bowman, S. and Zhu, X. (2019). Deep learning for natural language inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 6–8, Minneapolis, Minnesota. Association for Computational Linguistics.
- [Bowman et al. 2015] 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*.
- [Chen et al. 2018] Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., and Wei, S. (2018). Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.
- [Das and Chakraborty 2018] Das, B. and Chakraborty, S. (2018). An improved text sentiment classification model using tf-idf and next word negation. *arXiv preprint arXiv:1806.06407*.
- [Hartmann et al. 2017] Hartmann, N. S., Fonseca, E., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.
- [Howard and Ruder 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- [Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- [Oliveira da Silva et al. 2023] Oliveira da Silva, F., Craveiro, G. M., Siqueira Souza, J. M., Silva, V. F. d., and Vanzin, V. J. d. B. (2023). Natural language inference bow, word embeddings and symbolic experiments at assin-2 dataset. GitHub repository, https://github.com/jmssouza/nlp_entailment/.
- [Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- [Prokhorenkova et al. 2018] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.
- [Real et al. 2020] Real, L., Fonseca, E., and Gonalo Oliveira, H. (2020). *The ASSIN 2 Shared Task: A Quick Overview*, pages 406–412.
- [Schuster and Paliwal 1997] Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.
- [Shlens 2014] Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

- [Souza et al. 2020] 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.
- [Wagner Filho et al. 2018] 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)*.
- [Williams et al. 2018] Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- [Zhang et al. 2010] Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52.

Previsão de Utilidade de Avaliações de Produtos Online na Língua Portuguesa Brasileira

Larissa F. S. Britto^{1,2}, Luciano D. S. Pacífico³, Teresa B. Ludermir¹

¹Centro de Informática – CIn

Universidade Federal de Pernambuco – UFPE – Recife, PE – Brasil

²Centro de Pesquisa e Desenvolvimento em Telecomunicações –
CPQD – Campinas, SP – Brasil

³Departamento de Computação (DC)

Universidade Federal Rural de Pernambuco – UFRPE – Recife, PE – Brasil

lfsb@cin.ufpe.br, luciano.pacifico@ufrpe.br, tbl@cin.ufpe.br

Abstract. *With the growth of e-commerce, online product reviews have become a significant factor in influencing users' purchasing decisions. However, users may be harmed by the information overload on online review platforms. In this study, we evaluate different approaches to identify helpful product reviews. To achieve this, a large dataset of Amazon reviews from various product domains was proposed. The results demonstrate that it is possible to predict the usefulness of online reviews without relying on any handcrafted features.*

Resumo. *Com o crescimento do comércio eletrônico, as avaliações de produtos online se tornaram um fator importante na decisão de compra dos consumidores. No entanto, os usuários podem ser prejudicados pela sobrecarga de informações em plataformas de avaliação online. Neste estudo, avaliamos diferentes abordagens para identificar avaliações de produtos úteis. Para esse propósito, foi proposto um grande conjunto de dados de avaliações da Amazon em diferentes domínios de produtos. Os resultados mostram que é possível prever a utilidade das avaliações online sem depender de recursos personalizados sem depender de quaisquer características feitas manualmente.*

1. Introdução

Com o crescimento da Internet, o comércio eletrônico se tornou um dos métodos de compra mais importantes. O processo de tomada de decisão de compra é único para as compras online. Vários elementos motivacionais, como fatores situacionais, características do produto e experiências anteriores de compras online, podem influenciar as atitudes dos consumidores em relação às compras online [Senecal et al. 2005].

As avaliações de usuários têm um forte impacto na decisão de compra. Dados significativos sobre opiniões de usuários, é uma fonte rica de conhecimento para a Análise de Sentimentos (AS), área focada em detectar sentimentos e opiniões em textos [Henrickson et al. 2019, Tonkin 2016].

Uma tarefa de AS que ganhou popularidade é a predição de utilidade, que visa resolver o problema da sobrecarga de informações em plataformas de avaliação online,

que afeta a capacidade de clientes de avaliar a qualidade de produtos ou empresas ao tomar decisões de compra [Bilal and Almazroi 2022]. As plataformas de avaliação introduziram e implementaram um sistema de votos úteis (onde os usuários votam nas avaliações que consideram úteis), mas essa estratégia depende muito da cooperação do usuário.

Neste artigo, avaliamos experimentalmente diferentes abordagens estabelecidas na literatura de PLN para classificar a utilidade em avaliações de produtos online. Considerando que a grande maioria dessas abordagens se refere a corpora em inglês e que as aplicações em português brasileiro são relativamente escassas, propomos uma grande base de dados de avaliações de produtos da Amazon [Tufchi et al. 2023, ElKafrawy et al. 2023], para predição de utilidade.

O restante do artigo está organizado da seguinte forma. Na Seção 2, apresentamos e descrevemos a base de dados proposto. A Seção 3 descreve nossa configuração experimental. Os resultados experimentais e sua discussão são apresentados na Seção 4. A última seção, Seção 5, conclui o artigo.

2. Base de Dados

Nesta seção, a base de dados utilizado neste trabalho são apresentados e brevemente descritos.

2.1. Data Collection

A base de dados proposto foi extraído do site de vendas Amazon ¹, uma das plataformas mais populares para leitura e postagem de avaliações.

Nossa base de dados considera vários domínios de produtos. Para a coleta, selecionamos algumas das categorias mais populares na plataforma. Para cada uma dessas categorias, coletamos todas as avaliações em português do Brasil dos 100 produtos mais vendidos. Além das avaliações, também foram coletadas outras informações que podem ser aplicadas para outras tarefas de análise de sentimentos, como informações sobre o produto e o usuário. A coleta de dados foi realizada entre 28 de junho e 31 de junho de 2022. O Framework Scrapy² para extração de dados de websites foi adotado nesta etapa.

2.1.1. Processamento de Dados

Após a coleta de dados, realizamos uma etapa de processamento para normalizar os dados numéricos. Os dados textuais estão disponíveis sem nenhum pré-processamento, para que os pesquisadores que desejam usar esta base de dados possam escolher os métodos mais adequados de acordo com sua pesquisa. As estatísticas da base de dados estão listadas na Tabela 1.

2.2. Anotação dos Dados

A anotação de utilidade foi feita com base nas informações de votos úteis. Como a plataforma da Amazon não fornece informações sobre votos não úteis, neste trabalho, assumimos que qualquer avaliação sem pelo menos um voto útil é considerada não útil.

¹<https://www.amazon.com.br/>

²<https://scrapy.org/>

Table 1. Estatísticas da base de dados por classe. Símbolos como emojis não são considerados, resultando em algumas avaliações com comprimento 0.

Medida	Útil	Não Útil	Total
Tamanho Máximo	250	221	250
Tamanho Mínimo	0	0	0
Tamanho Médio	27,47	14,44	15,49
Tokens Únicos	28783	55308	61605
Número de Documentos	30516	348883	379399

2.3. Balanceamento dos Dados

Nossa base de dados apresenta um alto grau de desbalanceamento, o que poderia afetar o desempenho dos classificadores. Para balancear nossa base de dados, reduzimos o tamanho da classe que é abundante, removendo aleatoriamente documentos dessa classe.

2.4. Disponibilidade dos Dados

A base de dados final e o script para seu desenvolvimento serão disponibilizados no seguinte repositório público: www.github.com/larifelicianahelpful-Amazon-PT.

3. Configuração Experimental

A avaliação experimental deste trabalho tem como objetivo comparar o desempenho de modelos de classificação na predição de utilidade. Foi selecionado um método popular da literatura de classificação de texto para extração de características: TF-IDF. Essas características foram utilizadas como entrada para classificadores tradicionais de aprendizado de máquina: k-vizinhos mais próximos (k-Nearest Neighbors), regressão logística (Logistic Regression), Naive Bayes, Floresta Aleatória (Random Forest) e Máquinas de Vetores de Suporte (Support Vector Machines). Além desses, dois modelos BERT [ElKafrawy et al. 2023] foram ajustados (*fine-tuned*) e utilizados como modelos de classificação: BERTimbau [Souza et al. 2020] (modelos BERT pré-treinados para o português brasileiro) e BERT Multilíngue [Devlin et al. 2018] (pré-treinado em 104 idiomas, incluindo o português).

Na nossa avaliação, foi utilizada a validação cruzada com cinco *folds*, na qual a base de dados proposta foi dividida aleatoriamente em cinco partes balanceadas para formar o conjunto de treinamento e o conjunto de teste. Quatro partes são usadas cada vez para formar o conjunto de treinamento, e a parte restante é usada como conjunto de teste. O processo de reamostragem foi realizado para evitar resultados obtidos por acaso. Foram adotadas métricas de classificação bem conhecidas: Precisão Macro e Micro, Revocação (Recall) e Medida-F (F-measure).

4. Resultados e Discussão

Nesta seção, os resultados experimentais são apresentados e discutidos. A Tabela 2 mostra os resultados de todos os modelos na predição de utilidade.

Conforme mostrado nos resultados da Tabela 2, alguns classificadores tradicionais de aprendizado de máquina tiveram bom desempenho, como a Regressão Logística e a Floresta Aleatória. O SVM obteve o melhor desempenho, alcançando 84,3% de precisão. Os modelos BERT tiveram desempenho muito semelhante ao SVM, alcançando uma precisão de 84,1%. Apesar de terem alcançado um bom desempenho, esses os classificadores

Table 2. Resultados experimentais para predição de utilidade. Melhores resultados para cada métrica aparecem em negrito.

Modelo	Acurácia	Precisão	Revocação	F-Measure	Treinamento (s)	Teste (s)
TF-IDF + KNN	0.7975	0.8947	0.6746	0.7691	2.01	25.46
TF-IDF + LR	0.835	0.8521	0.8108	0.8309	4.45	0.49
TF-IDF + NB	0.8372	0.9027	0.7558	0.8227	2.02	0.49
TF-IDF + RF	0.8306	0.8637	0.7852	0.8225	103.68	1.33
TF-IDF + SVM	0.8427	0.8751	0.7995	0.8356	1067.68	111.92
BERT (PT-BR)	0.8409	0.8641	0.8091	0.8357	769.74	18.47
BERT (Multilingual)	0.8353	0.8790	0.7781	0.8253	869.06	18.70

apresentaram um longo tempo de execução médio para treinamento e teste, ambos levando mais de 13 minutos. Em contrapartida, os classificadores com desempenho médio, como a Regressão Logística, tiveram um tempo de execução baixo, levando apenas alguns segundos.

Os dois modelos BERT tiveram um desempenho semelhante entre si, o que poderia indicar que ambos os modelos podem ser uma boa opção para essa tarefa.

5. Conclusões

A quantidade de opiniões fornecidas pelos usuários na Internet todos os dias tem feito com que tarefas de análise de sentimentos sejam altamente requisitadas pelas empresas. Dois desses desafios são abordados neste artigo: predição de utilidade. Um conjunto de dados extenso e rico em informações foi desenvolvido por meio da coleta de avaliações de produtos da Amazon em português do Brasil.

Diferentes abordagens da literatura para análise de sentimentos foram adotadas para extração de características e classificação. O BERT, uma técnica de ponta para várias tarefas de Processamento de Linguagem Natural, é usado neste estudo juntamente com uma técnica tradicional de extração de características, TF-IDF, e classificadores de aprendizado de máquina (k-Nearest Neighbors, Logistic Regression, Naive Bayes, Random Forest e Support Vector Machines).

Os modelos BERT selecionados (BERTimbau e Multilingual), apesar do longo tempo de treinamento, demonstraram ser ótimos modelos para ambas as tarefas, alcançando alta precisão. O classificador SVM também obteve ótimos resultados, porém enfrentou problemas de escalabilidade devido ao longo tempo de treinamento e teste.

Existem algumas limitações associadas à nossa metodologia em termos de rotulação de ajuda. Rotular avaliações sem classificação de ajuda como *não útil* pode não ser a melhor abordagem, pois várias razões podem levar a uma avaliação não ser votada, como baixa demanda pelo produto e avaliações publicadas recentemente (com pouca visualização). Apesar desse problema, os resultados garantem a qualidade da base de dados e mostram como essa tarefa pode ser realizada sem o uso de características criadas manualmente.

Em trabalhos futuros, pretendemos estender a avaliação comparativa com outros modelos estado-da-arte da literatura de classificação de texto, como CNN e BiLSTM. Também pretendemos analisar como diferentes etapas de pré-processamento (como remoção de stopwords e stemming) podem afetar os resultados.

References

- Bilal, M. and Almazroi, A. A. (2022). Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews. *Electronic Commerce Research*.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- ElKafrawy, P., Mahgoub, A., Atef, H., Nasser, A., Yasser, M., Medhat, W. M., and Darweesh, M. S. (2023). Sentiment analysis: Amazon electronics reviews using bert and textblob.
- Henrickson, K., Rodrigues, F., and Pereira, F. C. (2019). Chapter 5 - Data Preparation. In Antoniou, C., Dimitriou, L., and Pereira, F., editors, *Mobility Patterns, Big Data and Transport Analytics*, pages 73–106. Elsevier.
- Senecal, S., Kalczynski, P., and Nantel, J. (2005). Consumers' decision-making process and their online shopping behavior: A clickstream analysis. *Journal of Business Research*, pages 1599–1608.
- 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*, page 403–417, Berlin, Heidelberg. Springer-Verlag.
- Tonkin, E. L. (2016). Chapter 2 - A Day at Work (with Text): A Brief Introduction. In Tonkin, E. L. and Tourte, G. J. L., editors, *Working with Text*, Chandos Information Professional Series, pages 23–60. Chandos Publishing.
- Tufchi, S., Yadav, A., Rai, V. K., and Banerjee, A. (2023). Sentiment analysis on amazon product review: A comparative study. In Khanna, A., Polkowski, Z., and Castillo, O., editors, *Proceedings of Data Analytics and Management*, pages 139–149, Singapore. Springer Nature Singapore.

Proceedings of 14th Symposium in Information and Human Language
Technology

JDP 2023

VIII Jornada de Descrição do Português

Tipologia de fenômenos ortográficos e lexicais em CGU: o caso dos *tweets* do mercado financeiro

Clarissa Lenina Scandarolli^{1,2}, Ariani Di Felippo^{1,2}, Norton Trevisan Roman^{1,3}
Thiago A. S. Pardo^{1,4}

¹Núcleo Interinstitucional de Linguística Computacional – NILC

²Departamento de Letras – Universidade Federal de São Carlos – UFSCar
Caixa Postal 676 – CEP 13565-905 – São Carlos – SP – Brasil

³Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
CEP 03828-000 – São Paulo – SP, Brasil

⁴Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13566-970 – São Carlos – SP – Brasil

clarissa.scandarolli@estudante.ufscar.br, arianidf@gmail.com,
norton@usp.br, taspardo@icmc.usp.br

Abstract. Twitter is an attractive source of information for several Natural Language Processing (NLP) applications, especially sentiment analysis and opinion mining. In this paper, we present a systematic description of orthographic and lexical phenomena in a corpus of tweets from the stock market domain in Portuguese. As a result, we propose a typology of the phenomena that could support the definition of annotation guidelines for their treatment within the Universal Dependencies framework of syntactic analysis and the development of NLP applications that realize term disambiguation or probabilistic ordering of options, as is the case with suggestions presented to users by spelling checkers.

Resumo. Twitter é uma fonte atrativa de informação para várias aplicações do Processamento Automático das Línguas Naturais (PLN), especialmente análise de sentimento e mineração de opinião. Neste artigo, apresenta-se uma descrição de fenômenos ortográficos e lexicais em um corpus de tweets do mercado financeiro em português. Como resultado, propõe-se uma tipologia dos fenômenos que pode auxiliar na definição de diretrizes de anotação segundo o modelo gramatical Universal Dependencies e no desenvolvimento de aplicações de PLN que façam a desambiguação de termos ou a ordenação probabilística de opções, como ocorre com a escolha das sugestões ortográficas apresentadas ao usuário em um corretor ortográfico.

1. Introdução

O *Twitter* é uma fonte de informações valiosas para diferentes segmentos da sociedade devido principalmente à influência dessas informações. Por conseguinte, aplicações linguístico-computacionais (p.ex.: análise de sentimento e mineração de opinião) que processam o conteúdo gerado pelos usuários (CGU) do *Twitter* têm sido muito desenvolvidas no Processamento Automático das Línguas Naturais (PLN) [Sanguinetti et al. 2022]. E esse desenvolvimento é desafiador devido à linguagem não-padronizada

dos *tweets*, que pode ter sentenças agramaticais, sequências de sintagmáticas curtas, palavras com ortografia não convencional e expressões específicas de domínio. Para o desenvolvimento das aplicações, já há etiquetadores morfossintáticos (*taggers*) e analisadores sintáticos (*parsers*). Tal ferramental, aliás, tem sido construído com base nos *treebanks* ou *corpora* anotados (comumente com informações morfossintáticas e sintáticas) [Sanguinetti et al. 2022]. Os *tweebanks* mais recentes possuem anotação segundo o modelo gramatical *Universal Dependencies* (UD) [Nivre et al. 2016].

Motivados pela necessidade de criação de diretrizes para a anotação-UD de *tweebanks*, autores como Sanguinetti et al. (2022) focaram em descrever as idiossincrasias linguísticas mais gerais dos CGUs, propondo uma tipologia. Isso porque, mesmo o CGU sendo um contínuo de subdomínios textuais que variam de acordo com (i) convenções e limitações específicas impostas pela plataforma utilizada (como blog, fórum de discussão, chat online, microblog, etc.), (ii) grau de “canonicidade” em relação a uma linguagem mais padronizada e (iii) dispositivos linguísticos adotados para transmitir uma mensagem, há fenômenos comuns a esse espectro. Embora haja fenômenos comuns aos diferentes tipos de CGUs, a linguagem pode ser fortemente marcada pelo domínio (ou assunto) do *corpus*.

Assim, apresenta-se aqui a descrição das características ortográficas/gráficas e lexicais do *corpus* DANTEStocks¹, que engloba 4.048 *tweets* em português sobre o mercado financeiro. Acredita-se que a tipologia resultante da sistematização dos fenômenos pode auxiliar no processo de normalização dos *tweets*, desenvolvimento de aplicações multigênero ou de uso geral que não requerem normalização e na definição de diretrizes de anotação segundo o modelo gramatical UD.

2. Trabalhos relacionados

Estudos sobre variantes ortográficas da língua padrão têm longa tradição no PLN, sobretudo devido às aplicações de correção ortográfica. Muitas das pesquisas se baseiam nas 4 categorias de desvios de Damerau (1964) (inserções, exclusões, substituições e transposições de letras). Com o objetivo de verificar se essas classes se aplicavam ao português, Gimenes et al (2014), por exemplo, investigaram um *corpus* de *blogs* de viagens e comentários e, além das 4 categorias de Damerau, identificaram 3 categorias extras: erros no uso de diacríticos, erros no uso da cedilha e erros relacionados à espaço.

Sobre os *tweets* e gêneros similares, Bertaglia (2017), por exemplo, visando à construção de ferramentas de normalização para UGC, investigou um *corpus* em português composto por *tweets*, postagens de um fórum de discussão e análises de produtos. O autor identificou 3.699 palavras distintas que não constavam em um dicionário de referência e anotou essas palavras em função de 8 categorias de desvios da língua padrão: (i) erro ortográfico (e de digitação), (ii) acrônimo, (iii) abreviação, (iv) internetês, (v) estrangeirismo, (vi) unidade de medida, (viii) nome próprio, e (vii) sem categoria (isto é, *tokens* cuja classificação não é clara ou varia conforme o contexto). Sanguinetti et al. (2022) propuseram uma sistematização das particularidades identificadas em *corpora* majoritariamente compostos por *tweets* com base em 2 dimensões: canonicidade e intencionalidade. Por “canonicidade”, entende-se a

¹ <https://drive.google.com/file/d/1wr9M4czkPgkUj1--U9GT9h8ncXc6rzv4/view?usp=sharing>

propriedade de um fenômeno ocorrer na língua padrão ou não. “Intencionalidade” se refere ao fato do fenômeno ter sido produzido deliberadamente ou não. Na hierarquia dos autores, “marcas de expressividade”, por exemplo, são um tipo de fenômeno não-canônico e intencional, com os subtipos: (i) reduplicação de pontuação (“!”→“!!!”), (ii) alongamento grafêmico (“linda”→“linnda”), (iii) *emoticons* (“:-)”) e (iv) *emojis* (“❤”).

A seguir, apresenta-se o *corpus* DANTEStocks², que foi alvo deste trabalho.

3. O *corpus* DANTEStocks

O DANTEStocks é um *corpus* de UGC em português composto por *tweets* sobre o mercado financeiro. Ele resultou do refinamento e da anotação morfossintática do *corpus* de Silva et al. (2020), cuja compilação se baseou na ocorrência de menos um *ticker*³ de uma das 73 ações do IBovespa (principal indicador de desempenho das ações negociadas na B3). Atualmente, o DANTEStocks possui 4.048 *tweets* (~81 mil *tokens*), os quais não foram submetidos nenhuma normalização e, por terem sido compilados em 2014, têm no máximo 140 caracteres. Quanto à estrutura, o *corpus* engloba *tweets* com diferentes constituições internas, podendo apresentar (i) uma ou mais sentenças bem delimitadas (1) e (2), (ii) ausência de pontuação (3) ou pontuação equivocada (4), (iii) fragmentação (5), e (iv) colagens de manchetes de outras fontes (6) [Di-Felippo et al. 2021].

- (1) Sera k petr4 já entrou na baixa?
- (2) PETR4 subiu na bolsa 13,50. Muito bem, surpreso com o resultado.
- (3) #PT conseguiu fazer propaganda eleitoral antecipada O que a @user⁴ tem a dizer sobre isso?
- (4) Bom dia Marcos, Alguma previsão para petr4?!
- (5) #GGBR4 Suportes e resistências <http://t.co/Azw6yIEVI9>
- (6) Logística, ex-LLX, anuncia prejuízo de R\$ 135,8 milhões em 2013: A Prumo Logística, ex-LLX (LLXL3), divu... <http://t.co/LwmlKPqssk>.

O DANTEStocks possui anotação de emoção, realizada manualmente com base nos 4 eixos de oposição emocional da teoria de *Plutchik* [Plutchik e Kellerman 1986] (*joy* vs *sadness*, *anger* vs *fear*, *trust* vs *disgust* e *surprise* vs *anticipation*) [Silva et al. 2020]. O *tweet* (1), por exemplo, recebeu os seguintes rótulos para 3 dos pares emocionais: *joy*, *trust* e *surprise*. O DANTEStocks também possui anotação semiautomática em nível morfológico segundo a UD, na qual se especificaram o lema, a etiqueta morfossintática e os traços lexicais/gramaticais (*features*) das palavras. O outro nível de anotação, no qual se explicitam as relações sintáticas de dependência (*deprels*), ainda não foi anotado. Na Figura 1, ilustra-se a anotação-UD completa de um *tweet* do *corpus* com base em Sanguinetti et al. (2022). Nessa figura, as etiquetas morfossintáticas (*part-of-speech* ou *PoS*)⁵ estão em caixa alta, como NOUN para “acordo”. Acima, estão os lemas, como “voo” para “voos”. As *deprels* estão indicadas por setas rotuladas que se originam no *head* e se destinam ao dependente. Na figura, “acordo” é dependente de

² <https://drive.google.com/file/d/1wr9M4czkPgkUj1--U9GT9h8ncXc6rztv4/view?usp=sharing>

³ Em (1), por exemplo, o *ticker* “petr4” indica ações preferenciais da Petrobras.

⁴ As menções aos usuários do *Twitter* foram anonimizadas.

⁵ A versão 2.0 da UD dispõe de 17 *tags* de *PoS* e de critérios para o emprego/anotação de cada uma delas.

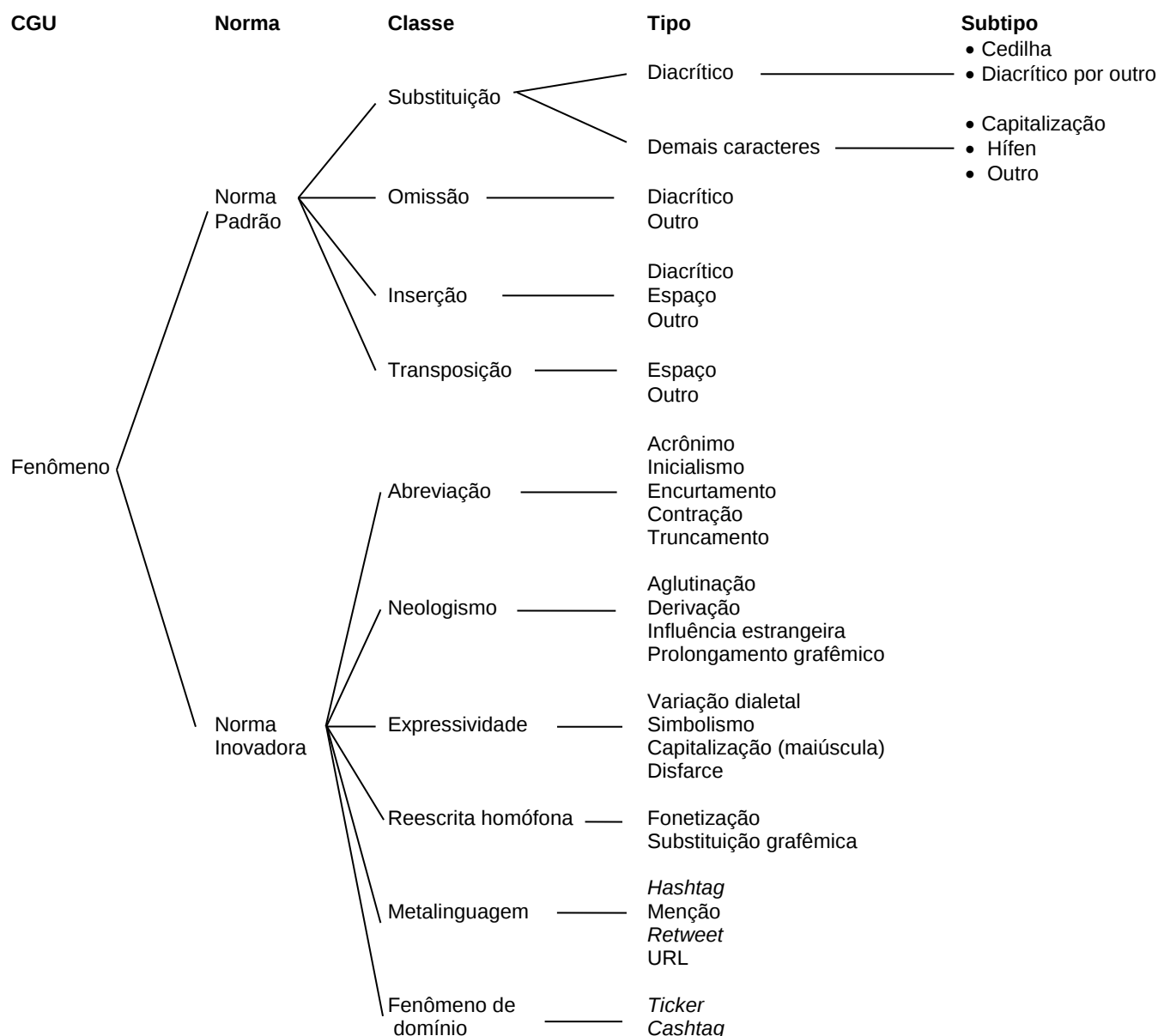


Figure 2. Tipologia de idiossincrasias gráficas/lexicais do DANTEStocks.

A “Norma Padrão” engloba fenômenos considerados desvios da norma-padrão por diversos motivos (como desconhecimento da ortografia, influência do meio e dispositivo, influência de novas regras fonéticas, etc.). As categorias dessa dimensão foram definidas com base no conceito de “caractere” do padrão Unicode^{9,10}. Nesse padrão, letras com diferentes capitalizações (“a” e “A”), diferentes acentuações (“o” e “ó”), diacríticos (em isolado) (“~”) e o próprio espaço são caracteres diferentes e, por isso, representados por códigos únicos. Para ilustrar, a letra minúscula “o” tem o código U+006F e a letra maiúscula “O” é codificada por U+004F. Além disso, um caractere como “á” pode ser concebido como a composição de 2 *code points* (p. ex.: ao se digitar “'” + “a”, obtém-se o “á”), o qual, nos algoritmos de normalização do Unicode, é

⁹ <http://www.unicode.org/standard/WhatIsUnicode.html>

¹⁰ Há *code points* para mais de 1 milhão de caracteres, permitindo que as máquinas representem e manipulem de forma consistente texto de qualquer sistema de escrita.

convertido para um *code point* único. Sendo assim, a adoção do caractere permite identificar e classificar as variações gráficas com base em um critério genérico, abrangente e concreto, além de amplamente empregado na Computação. Aplicando o conceito de caractere às categorias de Damerau, os fenômenos dessa dimensão foram organizados em classes, tipos e subtipos. Ressalta-se que um mesmo *token* pode apresentar mais de um fenômeno da Norma Padrão. A “Norma Inovadora” engloba fenômenos que são empregados de modo a concorrer com outras palavras da linguagem-padrão para expressar um mesmo conceito ou de modo a expressar um conceito novo. De certa forma, essa norma se relaciona às “variantes linguísticas” de uma comunidade de fala, as quais, necessariamente, não estão contempladas na norma-padrão, pois são resultados da utilização de recursos ortográficos de forma criativa e inovadora.

1) Norma Padrão

- *Substituição*: ocorre quando ao menos um caractere (diacrítico ou não) de um *token* é substituído por outro, ocasionando um erro da ortografia padrão. A substituição de diacrítico pode ser de 2 tipos: (i) cedilha, como “acougue” (“açougue”), pois, embora haja um *code point* único para o diacrítico do cedilha, este é indissociável da letra “c”, e (ii) diacrítico por outro, como “mâe” (“mãe”). Os demais casos de substituição envolvem (i) capitalização (maiúscula e minúscula), como “dilma” (“Dilma”), (ii) substituição de hífen por espaço (e vice-versa), como “cruz credo” ao invés de “cruz-credo”, ou (iii) outro caractere, como ocorre em “hirário” ao invés de “horário”.
- *Omissão*: ocorre quando um caractere deixa de ser expresso. Uma omissão pode ser relativa a (i) diacrítico, pois o usuário deveria ter digitado duas ou mais teclas para compor o caractere, mas não o fez, como “esta” (“está”), ou (ii) demais caracteres, como a ausência do *s* plural no final de “açõe” (“ações”).
- *Inserção*: ocorre quando um caractere é inserido da palavra. Uma inserção pode ser relativa a (i) diacrítico, como “Petrobrás”, quando o correto seria “Petrobras”, (ii) espaço, como “a final” ao invés de “afinal”, e (iii) outro caractere, como “Streaddle”, quando o correto é “Straddle”.
- *Transposição*: ocorre quando um caracter é trocado de ordem com outro. Uma transposição pode ser de (i) espaço, como “meua migo”, quando o correto é “meu amigo”, ou de (ii) demais caracteres, como “acrodo” ao invés de “acordo”.

2) Norma Inovadora

- *Abreviação*: fenômeno que gera um *token* mais curto do que a palavra ou expressão que lhe deu origem, podendo ser: (i) acrônimo, isto é, *token* composto pelas letras iniciais ou sílabas de uma multi-palavra e que tem pronúncia de palavra única, como “Cemig” (“Companhia Energética de Minas Gerais”); (ii) inicialismo, que se observa em um *token* composto pelas letras iniciais de uma multi-palavra e que é pronunciado letra por letra, como “lp” (“longo prazo”); (iii) encurtamento, isto é, ausência das letras finais de *token*, como “q” (“que”), (iv) contração, observado em um *token* com letras intermediárias ausentes, como “enqt” (“enquanto”), e (v) truncamento, isto é, *token* quebrado que, no caso do DANTEStocks, ocorre no final do *tweet*, comumente seguido por reticências, e que se deve ao limite de caracteres, como “divu” (“divulgou”) no exemplo (6).

- *Neologismo*: resulta em uma palavra nova ainda não institucionalizada (isto é, não abonada e incluída em dicionário), podendo ser de 3 tipos: (i) aglutinação, que se observa em um *token* resultante da junção de 2 palavras, como “Ibolixo” (“Ibovespa” + “lixo”); (ii) derivação, que resulta da adição de um afixo a uma radical já existente, como “diretassa” (“direta” + “-assa (-aça)”) e (iii) influência estrangeira, que se observa em uma palavra formada com base em outra língua, como “estopar”, que provém do verbo em inglês “*stop*” (“parar”) e significa “interromper venda ou compra de um ativo diante de dado preço”.
- *Expressividade*: fenômeno que majoritariamente simula sentimento expresso pela prosódia, expressão facial ou gesto na interação direta, podendo ser: (i) prolongamento grafêmico, como “noosaaa” (“nossa”), (ii) variação dialetal, como “malmita” (“marmita”), (iii) simbolismo, isto é, ocorrência de um caractere simbólico (seja *emoticon*, *emoji*, *smiley* ou outro) em substituição a uma palavra ou parte dela, (iv) capitalização, como “FEIO” no *tweet* (7) “#btow3: eita papel FEIO. #goll4 de olho na média móvel”, e (v) disfarce, que é a substituição de uma ou mais letras por um caractere especial para indicar autocensura, como “m*” (“merda”).
- *Reescrita homófona*: refere-se a uma variação gráfica motivada pela fonética ou pela simplificação de diacríticos, podendo ser: (i) fonetização, que é a representação da fala na escrita, como “krai” (“caralho”), e (ii) substituição grafêmica, que é o uso de uma letra a mais em substituição a um diacrítico, como “neh” (“né”) e “tou” (“tô”).
- *Metalinguagem*: corresponde a todo *token* que tipicamente ocorre no *Twitter* e que, por isso, não está previsto em dicionários, como (i) *hashtag* (pe.x.: #PT em (3)), (ii) menção, como se observa em (3), (iii) marca de *retweet* (RT), como no *tweet* (8) “Região 24,60 a 24,65 RT @Live_Trade: Fibr3 observo p/ compra”, e (iv) URL, que se observa nos *tweets* (5) e (6).
- *Fenômeno de domínio*: todo *token* que ocorre recorrentemente em *tweets* do mercado financeiro, como os *tickers* (p.ex.: em (1), (2) e (6)) e *cashtags* como no *tweet* (9) \$PETR3 - Petrobras (petr) - Comunicado <http://t.co/mHuCIyQmFi>.

5. Considerações finais

Atualmente, tem-se definido o conjunto de diretrizes e *tags* para a anotação dos fenômenos da Figura 2. Tendo em vista a adoção do modelo UD, as *tags* estão sendo propostas em inglês.

Para os fenômenos da dimensão denominada Norma Padrão, objetiva-se adotar a *tag* `Type=Yes` na coluna FEATS (destinada a atributos morfológicos) do formato CoNLL-U¹¹ como indicado pela própria UD e, na coluna MISC (reservada para demais informações e cujas *tags* podem ser definidas para um *treebank* específico), poder-se-á empregar uma *tag* adicional, como [`SNorm:standard norm`], com os seguintes valores possíveis: [`Sub=substitution`, `Om=omission`, `In=insertion`, `Tr=transposition`, `Ot=other`]. A coluna MISC pode conter ainda outra *tag* para esses fenômenos, como [`Type:type`], a qual especificaria o desvio. Essa etiqueta poderia ter os valores [`Ced=cedilla`, `Dia=diacritic`, `Cap=capitalization`, `Hif=hifen`, `Spc=space`, `Ot=other`].

Quanto a classe das abreviações (Norma Inovadora), objetiva-se seguir a diretriz geral da UD que prevê o emprego da *tag* `Abbr=Yes` na coluna FEATS. Ademais,

¹¹ <https://universaldependencies.org/format.html>.

pretende-se empregar, na coluna MISC, uma *tag* adicional como [INorm: innovative norm], cujos valores possíveis correspondem aos 5 tipos de abreviações, que são: [Acr=acronym, Init=initialism, Short=shortening, Cont=contraction, Trunc=truncation]. Para os demais fenômenos da Norma Inovadora, pretende-se anotá-los por meio da *tag* INorm na coluna MISC, cujos valores possíveis representam os 16 tipos de fenômenos, a saber: [Aggl=agglutination, Der=derivation, Fgn=foreign, Ext=graphemic stretching, Dial: dialectal variation, Sym: symbolism, Upp: uppercase, Dis: disguise, Fon: fonetization, Subst: graphemic substitution, Hasht: hashtag, Me: mention, Ret: retweet, URL: URL, Tic: ticker, Casht: cashtag].

Uma vez que o conjunto de etiquetas estiver de fato definido, o DANTEStocks será inteiramente anotado, revisando os casos iniciais que deram origem à tipologia, e identificando outros que porventura não estavam na lista inicial. Na sequência, pretende-se fazer um levantamento estatístico dos fenômenos/*tags*, gerando uma caracterização do domínio/*corpus*. Validar a taxonomia em outro *corpus* de *tweets* é uma possibilidade de trabalho futuro. Por fim, ressalta-se que a descrição dos fenômenos ora apresentada não só pode contribuir para a definição de diretrizes de anotação-UD, mas também para que aplicações de PLN possam levar em conta a distribuição desses fenômenos, seja de forma geral ou em algum gênero ou domínio específico, de modo a permitir, por exemplo, a desambiguação de termos, ou a ordenação probabilística de opções, como ocorre com a escolha das sugestões ortográficas apresentadas ao usuário em um corretor ortográfico (p.ex.: [Gimenes et al. 2015]).

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

- Bertaglia, T.F.C. (2017). Normalização textual de conteúdo gerado por usuário. Dissertação, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos.
- Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.
- 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 the Proceedings of the 7th Workshop on Portuguese Description (JDP), p. 335-343.
- Faraco, C. A. (2008). Norma culta brasileira: desatando alguns nós. SP: Parábola Editorial.
- Gimenes, P., Roman, N. T., Carvalho, A. M. B. R. (2015). Spelling error patterns in Brazilian Portuguese. *Computational Linguistics*, 41(1): 175–183.
- Luotolahti, J., et al. (2015). Towards universal web parsebanks. In the Proceedings of the 3rd Depling 2015, p. 211–220. Uppsala University.

- Nivre, J. et al. (2016). Universal Dependencies v1: a multilingual treebank collection. In the Proceedings of the 10th LREC, p.1659-66. Portorož. ELRA
- Plutchik R., Kellerman, H. (ed.) (1986) Emotion: theory, research and experience. NY: Acad. Press.
- Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A.T., Lynn, T., Rehbein, I. Ruppenhofer, J., Seddah, D., Zeldes, A. (2020). Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In the Proceedings of the 12th LREC. p. 5240-50. Marseille, France. ELRA
- Silva, F.J.V., Roman, N.T., Carvalho, A.M.B.R. (2020). Stock market tweets annotated with emotions. In *Corpora*, 15(3), p. 343-354. Online ISSN: 1755-1676.
- Straka, M. (2018) UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In the Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207, Brussels, Belgium. ACL.

Sinalizadores retórico-discursivos: revisitando a anotação RST no corpus CSTNews

Roana Rodrigues¹, Jackson Wilke da Cruz Souza², Paula Christina Figueira Cardoso³

¹Programa de Pós-Graduação em Letras - Universidade Federal de Sergipe (UFS),
São Cristóvão, SE - Brasil

²Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia (UFBA),
Salvador, BA - Brasil

³Departamento de Computação Aplicada - Universidade Federal de Lavras (UFLA),
Lavras, MG - Brasil

roana@academico.ufs.br, jackcruzsouza@gmail.com, paula.cardoso@ufla.br

Abstract. *Rhetorical Structure Theory (RST) is a discourse theory in which the coherence of a text can be characterized by a tree structure, where the discourse units are the leaves and the nodes represent the rhetorical relations between them. Although it is known that the identification of connectives that indicate these relations plays an important role in text processing, the absence of a prototypical discourse marker does not eliminate the possibility of their interpretation. In this paper, we describe the analysis of a sample from a corpus already annotated with RST, aiming to identify how these relations are signaled in the discourse. The results highlight the importance of investigating other flags in addition to DMs.*

Resumo. *Rhetorical Structure Theory (RST) é uma teoria discursiva na qual a coerência de um texto pode ser caracterizada por uma estrutura de árvore, em que as unidades discursivas são as folhas e os nós representam as relações retóricas entre elas. Embora seja conhecido que a identificação de conectivos que indicam as relações desempenha um papel importante no processamento do texto, a ausência de um marcador discursivo (MD) prototípico não impede a possibilidade de sua interpretação. Nesta proposta preliminar, descreve-se a análise de um recorte de um corpus já anotado com RST, com o objetivo identificar como as relações são sinalizadas no discurso. Os resultados destacam a importância de investigar outros sinalizadores para além de MDs.*

1. Introdução

O modelo teórico RST (*Rhetorical Structure Theory*) é uma teoria linguístico-descritiva que tem o discurso como ponto de partida e visa analisar e descrever “fenômenos de ordem sintática, semântica e pragmática que se “gramaticalizam” nos textos” [Hirata-Vale e Oliveira 2014, p. 406]. Trata-se de um modelo muito utilizado no Processamento de Língua Natural (PLN), contribuindo sobretudo com a construção de *parsers* e de ferramentas automáticas de sumarização, tradução e avaliação de textos.

Tendo origem no trabalho de Mann e Thompson (1988), a RST se enquadra no chamado *Funcionalismo da Costa-Oeste Norte-Americana* e estabelece diretrizes para a anotação e descrição das relações retóricas (também nomeadas *discursivas* ou *de*

coerência) de um texto, ou seja, das relações que permitem a coesão e coerência textuais. Tem-se como objeto de estudo as relações estabelecidas entre *núcleo* e *satélite* - além da consideração de relações *multinucleares* -, explicitando-se a intenção (produção do falante/escritor) e o efeito (recepção do ouvinte/leitor) do que é dito. Na Figura 1, exemplificamos as relações retóricas de um fragmento de texto extraído do corpus CSTNews [Cardoso *et al.*, 2011]¹:

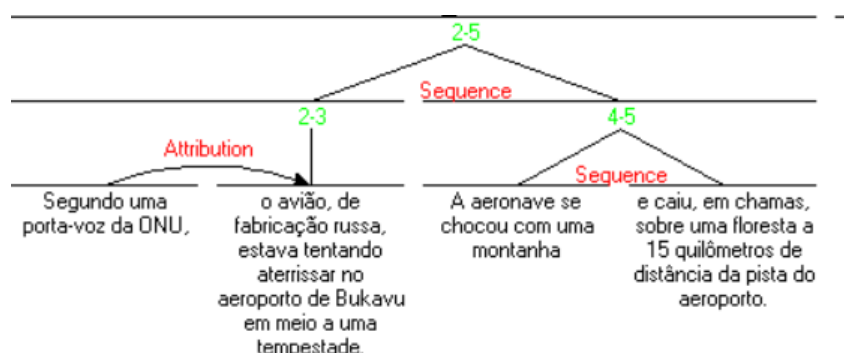


Figura 1. Exemplo de relações RST do corpus CSTNews

Na Figura 1, a árvore apresenta três relações retóricas: em 2-3, o *satélite* (*Segundo uma porta-voz da ONU*) caracteriza-se por apresentar a fonte de uma mensagem e o *núcleo*, a mensagem; tanto em 4-5, como em 2-5, têm-se relações multinucleares, pois apresentam núcleos em sequência, tendo como efeito o reconhecimento de uma sucessão temporal dos eventos.

Segundo Taboada e Mann (2006), a taxonomia adotada na RST não é fixa, isto é, não existe uma prescrição da teoria para o conjunto de relações retóricas possíveis para uma língua. No entanto, os autores salientam a necessidade de prudência na quantidade de relações, devido às dificuldades no processo manual de identificação e anotação de um texto. Para o português brasileiro (PB), Pardo (2005) propõe um conjunto de 32 relações retóricas possíveis², em que são estabelecidas informações relativas às restrições sobre o núcleo, o satélite, a relação núcleo-satélite, além de possíveis efeitos desencadeados no leitor.

As relações retóricas são comumente determinadas com base nos marcadores discursivos (MDs)³ presentes em um texto. Da anotação do CorpusTCC, Pardo (2005, p. 64-67) criou um quadro com a distribuição dos MDs em função das relações que

¹ A árvore em (1) foi visualizada e extraída na RSTTool, ferramenta utilizada para elaborar e abrir diagramas, disponível para *download* em: <http://www.wagsoft.com/RSTTool/section2.html>. Acesso em junho de 2023.

²Relações retóricas do PB, segundo Pardo (2005): *antithesis, attribution, background, circumstance, comparison, concession, conclusion, condition, contrast, elaboration, enablement, evaluation, evidence, explanation, interpretation, join, justify, list, means, motivation, non-volitional cause, non-volitional result, otherwise, parenthetical, purpose, restatement, same-unit, sequence, solutionhood, summary, volitional cause, volitional result*.

³ Assim como definido por Das e Taboada (2018), nesta pesquisa *marcadores discursivos*, também nomeados *conectivos*, são os elementos de um texto que estabelecem relações entre as proposições, incluem as conjunções, locuções conjuntivas, locuções preposicionais e expressões lexicalizadas.

sinalizam. Segundo o autor, embora nem todas as relações possuam MDs associados, os textos anotados com a relação *Sequence*, por exemplo, possuíam, majoritariamente, os marcadores *e, a partir de, em seguida*; já a relação *Explanation* foi marcada por *pois, isto é e porque*.

Taboada e Das (2013) destacam que a compreensão de textos parte da construção de uma representação das informações presentes nele, em que uma parcela desse processo compreende remontar as possibilidades de organização das proposições. Os autores apontam que a identificação de conectivos que indicam as relações possíveis facilita o processamento do texto, mas que a ausência de um marcador discursivo prototípico não furta a possibilidade de sua interpretação. Assim, argumentam que as relações de coerência são, na verdade, entidades cognitivas e, por conta disso, há possibilidade de interpretação do texto. Como resultado, trabalhos nessa perspectiva não consideram as relações não sinalizadas. Em (1), exemplo extraído de Das e Taboada (2018), as sentenças podem ser anotadas no modelo RST como *Contrast*; porém, seria uma relação implícita, já que não apresenta nenhum MD explícito. Apesar disso, é possível compreender o contraste quando considera-se as unidades lexicais *tall* e *short*.

(1) *John is tall. Mary is short.*⁴

Nesse sentido, em estudos mais recentes [Antonio 2017; Das e Taboada 2018] discute-se a necessidade de criação de tipologias dos sinalizadores discursivos para além dos marcadores, tais como *entonação, cadeia lexical, pontuação, tempo verbal*, entre outros, já que nem toda relação possui um marcador discursivo “explícito”/“prototípico” a ele relacionado.

Sendo assim, baseando-nos na proposta de Das e Taboada (2018), objetivamos analisar um recorte de um corpus já anotado em RST, para avaliar os sinalizadores presentes nos textos. Trata-se de uma caracterização preliminar da qual derivará uma tipologia de sinalizadores (simples e combinados) para as relações retóricas do PB. Para tanto, este artigo está organizado da seguinte maneira, além desta introdução: na seção 2, descrevemos os trabalhos de base para a presente investigação, ou seja, as contribuições de Antonio (2017) para o PB e Das e Taboada (2018) para o inglês; em seguida, apresentamos os processos e decisões metodológicas desta pesquisa; nossas primeiras percepções sobre o processo de anotação dos sinalizadores em corpus do PB e as considerações iniciais desta tarefa; e, então, as considerações finais e trabalhos futuros.

2. Sinalizadores discursivos

Nesta seção destacamos dois trabalhos relacionados aos objetivos traçados para esta pesquisa. É importante salientar que os trabalhos apresentados se baseiam em diferentes registros linguísticos: Antonio (2017) parte de um corpus oral, semi-formal e analisando o PB; já Das e Taboada (2018) analisam um corpus de textos jornalísticos do Inglês.

De acordo com Antonio (2017, p. 105), “as relações de coerência, por serem de sentido, e não de forma, podem ser estabelecidas e interpretadas independentemente de serem marcadas explicitamente por conectivos”. Posto isso, o autor investiga a

⁴ Tradução livre: “John é alto. Maria é baixa.”.

percepção de professores universitários em relações retóricas de 10 excertos de textos orais, considerando-se elementos para além dos MDs.

Como resultado, Antonio (2017) elenca as seguintes pistas formais destacadas pelos informantes: sinais de pontuação (dois pontos); modo de oração (*pergunta-resposta*); aspectos fonológicos (*entonação*); aspectos morfosintáticos (*tempo verbal, expressões adverbiais*); aspectos semânticos (*interdependência entre os estados-de-coisas; o próprio sentido das porções textuais envolvidas ou de palavras-chave nessas porções textuais, como paralelismo nas construções; referência anafórica*); e aspectos cognitivos (*ativação de referentes a partir de um modelo cognitivo global*).

Para o inglês, Das e Taboada (2018), a partir de um corpus já anotado com as relações retóricas (*RST Discourse Treebank*), realizam uma anotação minuciosa dos sinalizadores discursivos dessas relações, construindo assim o *RST Signalling Corpus*. Para tanto, os autores consideraram elementos formais para além dos MDs, organizando a taxonomia conforme se apresenta na Figura 2, retirado de Das e Taboada (2018).

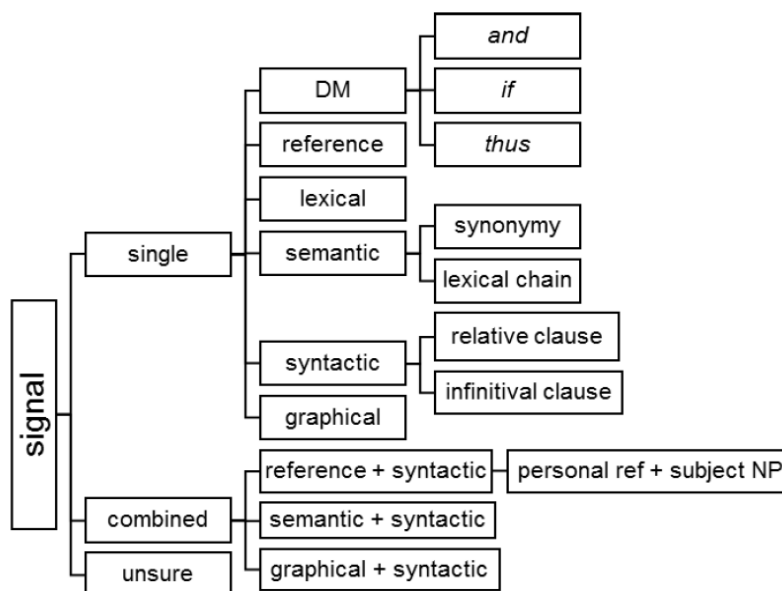


Figura 2. Fragmento da taxonomia hierárquica de sinalizadores discursivos

Na Figura 2, Das e Taboada (2018) pontuam que uma determinada relação pode ser anotada por um *sinalizador único* (tais como MDs, referência pessoal, oração relativa, dois pontos) ou um *sinalizador combinado* (vírgula + oração no particípio passado; construção sintática paralela + cadeia lexical, entre outros). Além disso, no processo de anotação, os autores relatam que houve casos anotados como *incertos*, nos quais não foi possível estabelecer com precisão o sinalizador que denota determinada relação.

3. Processos metodológicos

Para viabilizar esta investigação, foi selecionado o corpus CSTNews [Cardoso *et al.* 2011] que é anotado manualmente de diferentes maneiras quanto à organização do

discurso, sendo uma delas a RST. O corpus possui 50 conjuntos de textos (*clusters*), organizados por assunto, que foram coletados manualmente no ano de 2007. No total, são 140 textos jornalísticos, que juntos contabilizam 2.088 sentenças e 47.240 palavras.⁵

Os anotadores realizaram uma análise incremental, aproveitando a estrutura organizacional do texto fonte. Nessa abordagem, assume-se que as orações adjacentes dentro das sentenças devem ser relacionadas em primeiro lugar, seguido pelas sentenças adjacentes dentro dos parágrafos, e, por fim, os parágrafos adjacentes são relacionados. Na maioria das anotações, estabeleceram-se relacionamentos binários, ou seja, conectaram-se dois segmentos em uma relação. Ao final, observou-se a prevalência das relações *Elaboration*, *List*, *Attribution*, *Parenthetical* e *Same-unit* no corpus.

Nesta pesquisa partimos de um corpus anotado com RST, como proposto por Das e Taboada (2018) para a língua inglesa. Em nosso caso, utilizamos o CSTNews e selecionamos aleatoriamente, para a anotação manual dos sinalizadores, 9 *clusters*, que se constituem pelo conjunto de textos (de 2 a 3 textos jornalísticos) sobre a mesma notícia, totalizando 21 documentos anotados. Esse processo foi realizado por três anotadores em grupo, o que possibilitou discussões e tomadas de decisões conjuntas e imediatas. Para esta análise preliminar, decidimos identificar apenas sinalizadores intrasentenciais; os possíveis sinalizadores intersentenciais serão analisados numa fase posterior devido à dificuldade de haver consenso entre os anotadores, mesmo em um processo manual.

Com a anotação finalizada, passamos para a etapa de reflexão e análise dos sinalizadores apontados e as possíveis vinculações com relações retóricas específicas, conforme descrevemos na próxima seção.

4. Discussões e resultados

Antes da apresentação dos dados anotados, destacamos o fato de que, devido à decisão metodológica de anotação de documentos pertencentes a um mesmo *cluster*, muitas relações e sinalizadores se repetiram, por se tratar de proposições redundantes ou com pequenas e sutis variações. Além disso, indagamos a segmentação e anotação de algumas proposições do corpus, no entanto, neste trabalho preliminar, decidimos ignorar esses casos e investigá-los em trabalhos futuros.

Sendo assim, as discussões e resultados apresentados nesta seção são de cunho qualitativo, a partir de observações e discussões gerais dos dados anotados, com foco nas relações mais eminentes, a saber: *Attribution*, *Elaboration*, *Parenthetical*, *List*, *Sequence* e *Same-unit*. O Quadro 1 apresenta o tipo de sinalizador e as tags anotadas no corpus CSTNews para as relações destacadas.

⁵ Detalhes sobre o processo de anotação RST do corpus CSTNews, tais como anotadores, regras de segmentação e concordância da tarefa, podem ser encontrados em Cardoso *et al.* (2011).

Quadro 1. Relações retóricas e sinalizadores discursivos

Relação	Sinalizadores	Exemplos
<i>Attribution</i>	pontuação (aspas, vírgula) + informação sintática (verbo de comunicação (acrescentar, dizer, informar) e fonte). MD (que) + informação sintática (verbo de comunicação (acrescentar, dizer, informar) e fonte).	(2) ["Estamos resistindo à tentação de chamar o par de planeta duplo porque ele provavelmente não se formou do jeito que os planetas no nosso sistema solar apareceram"], [acrescenta Ivanov.] (3) [O ministro da Defesa, Nelson Jobim , informou no fim da noite desta terça-feira] [que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac)].
<i>Elaboration</i>	pontuação (vírgula) + pronome relativo (que). pontuação + passiva	(4) [Inicialmente, Solange Vieira,] [que é assessora especial de Jobim,] (5) [Segundo o jornal "Choson Sinbo"], [publicado pela Associação de Residentes Coreanos no Japão]
<i>List</i>	paralelismo + MD (e)	(6) (...) uma dupla de planetas errantes (...) [que giram ao redor deles mesmos] [e que vagam livremente pelo espaço.]
<i>Parenthetical</i>	pontuação (parênteses, travessão, vírgula) + sigla pontuação + mudança de tópico	(7) [O presidente do Conselho de Ética do Senado, Leomar Quintanilha] [(PMDB-TO)] (8) "(...) [publicado pela Associação de Residentes coreanos no Japão] [(próxima ao regime comunista da Coreia do Norte),]
<i>Same-unit</i>	pontuação (vírgula) + (concordância verbal (venceu)) + sucede uma relação encaixada (como <i>Elaboration</i> ou <i>Parenthetical</i>)	(9) [A seleção brasileira masculina de vôlei, que é treinada por Bernardinho ,] [venceu a Finlândia por 3 sets a 0.]
<i>Sequence</i>	MD (e) + tempo verbal pontuação (vírgula, ponto final) + numeral	(10) [Alvo de críticas incisivas da oposição desde o acidente com o Airbus da TAM, o atual presidente da Anac, Milton Zuanazzi, já teria concordado em renunciar] [e deve entregar o cargo nos próximos dias.] (11) [O time comandado pelo treinador Bernardinho só encontrou um pouco mais de dificuldades no segundo set.] [No terceiro , mesmo com vários reservas como o levantador Marcelinho e Samuel, os brasileiros conseguiram fechar a partida com tranquilidade.]

Conforme se observa no Quadro 1, as relações que se sobressaíram na anotação possuem sinalizadores combinados, gramaticalizados por MDs prototípicos somados a outros sinais, sobretudo *pontuação*, aspectos morfológicos (*tempo verbal*) e informações sintáticas (*construções de comunicação*, *passiva lexical*, *concordância verbal*, *orações subordinadas*).

A relação *Attribution* apresentou duas possibilidades de combinações de sinalizadores, relacionadas ao discurso em estilo direto (2) e indireto (3). Para a anotação manual, foi possível identificar a introdução do estilo indireto pela conjunção

(MD) *que*. No entanto, é sabido que, para qualquer anotação morfossintática, a multifuncionalidade de *que* é uma questão complexa para o processamento automático da língua. É o caso, por exemplo, da diferença encontrada entre *que* nas relações de *Attribution* e *que* nas relações de *Elaboration*, em que o primeiro é anotado como MD, mas o segundo atua como pronome relativo. Essa complexidade inerente à anotação morfossintática ressalta a necessidade de sinalizadores combinados para a identificação adequada das relações em análise.

As relações *Elaboration* e *Parenthetical* aparecem encaixadas a proposições nucleares, unidas sobretudo por sinais de pontuação (vírgula, travessão e parênteses). A relação *Parenthetical* se distingue ao ser anotada sempre em que são inseridas siglas no texto, além dos casos de *mudança de tópico*, característicos do acréscimo de uma informação adicional (colocada entre parênteses ou travessão). Por sua vez, a relação *Elaboration*, mais abundante no corpus, se assemelha ao comportamento das orações relativas (restritivas e explicativas), tendendo a ser introduzidas pelo pronome relativo *que*, embora apresente outros comportamentos, como se verifica em (5). Referente à *tag* [pontuação + passiva], salientamos a dificuldade de anotação para a distinção entre *particípio passado* (daí a justificativa para anotação da relação pela existência de uma *passiva lexical*) e *adjetivo*, visto que algumas relações de *Elaboration* verificadas no corpus se caracterizam pelo encaixe introduzido por um adjetivo, como em: [*Invicto na competição,*] [*o Brasil está tranquilo na liderança do Grupo B.*]. Portanto, em trabalhos futuros a anotação [passiva] deverá ser estudada com mais detalhes.

As relações *List* e *Sequence* são multinucleares, mas a primeira é identificada por relacionar itens comparáveis apresentados nos núcleos, enquanto a segunda se caracteriza por desencadear no leitor o efeito de reconhecimento de sucessão temporal dos eventos apresentados. Essa diferença de *restrição* e *efeito* se materializa nas *tags* anotadas para cada relação: apesar de ambas serem marcadas pela conjunção *e*, a relação *List* enfatiza a igualdade e comparação entre as proposições, indicada, em muitos casos, pela simetria entre as estruturas sintáticas (paralelismo), como em (6); já a relação *Sequence* é marcada pela ideia de sucessão, que se faz evidente tanto pela distinção temporal a partir da conjugação verbal (10), quanto pela ordem numérica de um determinado processo/progresso (11).

Por fim, destacamos os sinalizadores combinados da relação *Same-unit*, em que as informações apresentadas constituem uma única proposição. Na maioria dos casos, essa relação foi identificada devido à concordância verbal. No entanto, uma característica comum aos casos de *Same-unit* foi o fato de essa relação ser precedida por alguma relação RST encaixada (*Parenthetical* ou *Elaboration*), o que pode ser uma informação útil para o estabelecimento de regras para identificação automática dessa e de outras relações associadas.

Evidentemente, outras relações foram anotadas com sinalizadores simples e combinados no recorte do corpus anotado, mas propusemos a descrição dessas seis relações devido à sua frequência elevada. Os exemplos do Quadro 1, de (2) a (11), ilustram a importância de se reconhecer sinalizadores para além dos MDs e indicam caminhos, ainda que preliminares, sobre as características e a possibilidade de

identificação e reconhecimento automático das relações retóricas de um texto - ao menos do gênero jornalístico como o trabalhado nesta pesquisa.

5. Considerações finais e trabalhos futuros

Neste trabalho preliminar, nosso objetivo foi investigar pistas que pudessem sinalizar as relações do modelo teórico RST, partindo de um corpus do PB pré-anotado. O tipo de estudo exploratório que realizamos aqui demonstra a importância e a dificuldade em classificar relações RST considerando apenas MDs.

A ampla utilização de MDs na identificação de relações RST pode ser justificada por conta da possível compreensão das relações do modelo como *unidades de coerência discursiva*; nesse caso, seriam necessários conectivos específicos entre as unidades. Ainda nesse sentido, outra possível justificativa para essa utilização é poder compreender a RST como um modelo gramatical e, por conta disso, parece pertinente partir de MDs para caracterizar as relações do modelo. Entretanto, como apresentado, estudos recentes que utilizaram corpus de outros gêneros textuais (como de redes sociais) salientam a necessidade de explorar outros sinalizadores para além dos marcadores prototípicos.

Ademais, os resultados aqui apresentados apontam para a importância de não considerar MDs de maneira unívoca e como características exclusivas de algumas relações RST. Antes, destacamos a necessidade de analisar combinações entre os MDs e outros sinalizadores. Os marcadores “caso” e “eventualmente”, por exemplo, caracterizam a relação *Condition*, dado que não ocorrem em outras relações da teoria. Porém, identificamos que o marcador “mas”, característico da relação *Contrast*, foi utilizado em EDUs anotados com a relação *Concession*. Observamos comportamento similar com a conjunção “e”, que pode sinalizar tanto a relação *List* quanto a relação *Sequence*.

Quanto às limitações, destacamos o formato da anotação RST disponibilizado no corpus CSTNews. Iniciamos o trabalho identificando os possíveis sinalizadores com *tags* xml. Porém, há relações em que um único sinalizador pode estar entre informações que não foram consideradas na análise. A título de exemplo, tem-se a relação *Parenthetical*, em que a pontuação utilizada para identificar a relação, como travessões e parênteses. Decidir se anotamos com *tags* xml apenas os parênteses separadamente ou se anotamos incluindo o conteúdo dentre eles gera impactos diretos e substanciais na forma com que essa anotação será utilizada em classificadores automáticos, numa fase posterior a este estudo.

Quanto aos trabalhos futuros, pretendemos estender o estudo a outras porções textuais do corpus analisado, já que o estudo teve como ponto de partida a descrição de sinalizadores intrasentenciais. Outra tarefa a ser realizada, em estudo futuro, é ampliar a variabilidade do gênero textual, uma vez que o corpus utilizado é composto apenas por textos jornalísticos, garantindo que os sinalizadores que serão identificados possam ser provenientes de diferentes normas linguístico-gramaticais.

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, 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.
- Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (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, Brasil.
- 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.
- Hirata-Vale, F. B. M. e Oliveira, T. P. (2014) Modelos e Métodos de Análise Funcionalista. In: GONÇALVES, A. V.; GÓIS, M. L. S. (Org.). *Ciências da Linguagem: O Fazer Científico - Volume 2*. Campinas: Mercado de Letras.
- 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*, Vol. 8, N.3, pp. 243-281.
- Pardo, T. A. S. (2015) Métodos para análise discursiva automática. Tese (Doutorado em Ciências da Computação e Matemática Computacional). São Carlos: Universidade de São Paulo, 211p.
- Taboada, M. e Mann, W. C. (2006) Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, Vol. 8, N. 3, pp. 423-459.
- Taboada, M. e Das, D.. (2013) Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, V. 4, N. 2, pp. 249-281.

Complexidade textual em narrativas orais produzidas por informantes de diferentes níveis de escolaridade

Juliano Desiderato Antonio¹

¹Departamento de Teorias Linguísticas e Literárias – Universidade Estadual de Maringá
(UEM)

Maringá – PR – Brasil

jdantonio@uem.br

Abstract. *In this paper, a corpus formed by twenty oral narratives (ten produced by elementary school students and ten produced by undergraduate students) is analyzed with the aim of verifying whether the textual complexity of the narratives increases as the level of education increases. The tool used to analyze the textual complexity of the narratives in the corpus is the computational system NILC-Metrix, which employs two hundred metrics for this purpose. Eleven metrics were chosen which demonstrate that the narratives produced by the undergraduate students present higher textual complexity than the narratives produced by elementary school students.*

Resumo. *Neste trabalho, analisa-se um cópulo formado por vinte narrativas orais (dez produzidas por estudantes do ensino fundamental e dez produzidas por estudantes de curso superior) com o objetivo de se verificar se a complexidade textual das narrativas aumenta conforme aumenta o nível de escolaridade. A ferramenta utilizada para analisar a complexidade textual das narrativas do cópulo é o sistema computacional NILC-Metrix, que emprega duzentas métricas para esse fim. Escolheram-se onze métricas que demonstram que as narrativas do cópulo produzidas pelos alunos de curso superior apresentam maior complexidade textual do que as narrativas produzidas pelos alunos de ensino fundamental.*

Considerações iniciais

Com as recentes evoluções no campo do Processamento de Linguagem Natural (PLN), cada vez mais tarefas linguisticamente complexas vêm sendo realizadas por ferramentas computacionais. Alguns exemplos são o reconhecimento da fala humana, a tradução automática, a sumarização automática, a análise de sentimentos, dentre outros. Uma outra possibilidade interessante que vem se desenvolvendo, segundo Santucci et al. (2020), é a possibilidade de se analisar automaticamente a complexidade de textos do ponto de vista linguístico. Branco et al. (2014a), por exemplo, apresentam um sistema

que classifica automaticamente textos produzidos em Língua Portuguesa com base no Quadro Europeu Comum de Referência para Línguas. A classificação tem cinco níveis de dificuldade: A1 (mais fácil), A2, B1, B2 e C1 (mais difícil) e utiliza os critérios do Instituto Camões para certificação de proficiência. Os critérios utilizados para classificação, segundo Branco et al. (2014b) são leiturabilidade, densidade lexical, quantidade de sílabas por palavra, quantidade de palavras por sentença. Ao reduzirem um grande número de propriedades textuais a um número menor de dimensões de complexidade textual [Goldman e Lee 2014], as ferramentas podem também auxiliar na seleção de textos adequados para diferentes níveis de aprendizagem [Sheehan, Flor e Napolitano 2013; McNamara et al. 2014]. Evers (2018) descreve padrões léxico-sintáticos de redações submetidas ao vestibular da Universidade Federal do Rio Grande do Sul. A pesquisadora utiliza recursos e ferramentas dos Estudos do Léxico, da Linguística Textual, da Linguística de Córpus e da Linguística Computacional para identificar padrões lexicais e sintáticos correspondentes a três faixas de desempenho. Uma das ferramentas utilizadas pela pesquisadora trata especificamente da complexidade textual.

Jensen (2009) alerta para o fato de que dificuldade e complexidade textual são dois conceitos distintos que não devem ser confundidos. Enquanto a dificuldade é subjetiva e pode variar de leitor para leitor, a complexidade é mais objetiva e pode ser calculada a partir de critérios factuais, como os índices de leiturabilidade, a frequência das palavras (quanto mais comum é uma palavra menos esforço cognitivo é dispensado no processamento daquela palavra), a não literalidade (metáforas, metonímias e expressões idiomáticas podem afetar a complexidade textual pelo fato de exigirem maior esforço cognitivo para serem processadas).

Neste trabalho, investiga-se a complexidade textual de narrativas orais produzidas por informantes com diferentes níveis de escolaridade com a finalidade de se verificar se a complexidade textual das narrativas aumenta conforme aumenta o nível de escolaridade. O córpus é composto por dez narrativas produzidas por alunos do sexto ano do ensino fundamental e por dez narrativas produzidas por alunos de curso superior (Comunicação Social). A descrição dessas possíveis diferenças pode fornecer subsídios para que se descrevam os recursos empregados nos textos mais complexos. E esses recursos poderão ser utilizados em sala de aula por professores para auxiliarem seus alunos a produzirem textos narrativos mais complexos.

Metodologia

Quando da coleta dos dados, adotaram-se alguns critérios para que se evitasse ao máximo o risco de diferenças nos resultados causados por discrepâncias no córpus. Para que os textos de todos os informantes fossem sobre um mesmo assunto e fossem semelhantes em aspectos como extensão, conteúdo, etc, decidiu-se que a coleta dos

dados seria feita a partir da exibição de um vídeo com uma história que seria recontada pelos sujeitos da pesquisa. A opção pela narrativa proveio do fato de que, para a produção desse tipo de texto, o filme serviria como um *script* a ser seguido pelos informantes, o que permitiria a obtenção de um corpus bastante homogêneo. Para se evitar que houvesse influência das falas do narrador ou de personagens sobre a maneira como os informantes formulariam linguisticamente a história, a solução foi procurar um filme mudo, cuja sequência de cenas fosse suficiente para a compreensão do enredo. O vídeo escolhido foi “O pavão misterioso”, que se baseia em uma história do folclore nordestino de mesmo nome e que tem como personagens bonecos que representam seres humanos. Logo após assistirem ao filme, os informantes contaram a história oralmente, que foi gravada em fitas K-7.

A primeira parte do corpus foi coletada em 1996, com alunos do primeiro ano do curso de Comunicação Social de uma universidade situada no Norte do Paraná [Antonio 1998]. As demais narrativas foram coletadas em 2001, em uma escola estadual também situada em um município do Norte do Paraná. Os informantes eram alunos do sexto ano do ensino fundamental [Antonio 2004].

A ferramenta utilizada para analisar a complexidade textual das narrativas do corpus é o sistema computacional NILC-Metrix [Leal et al. 2022]. Esse sistema utiliza duzentas métricas propostas em estudos de Linguística Textual, Psicolinguística, Linguística Cognitiva e Linguística Computacional para investigar a complexidade textual no português brasileiro. Pode ser utilizado tanto com textos orais quanto com textos escritos.

As duzentas métricas são agrupadas em quatorze categorias [Leal et al. 2022]: índices descritivos, métricas de simplicidade textual, coesão referencial, coesão semântica, medidas psicolinguísticas, diversidade lexical, conectivos, léxico temporal, complexidade sintática, densidade do padrão sintático, informação morfossintática das palavras, frequência de palavras, fórmulas de legibilidade.

Por motivo de limitação de espaço, serão analisadas apenas onze métricas, as quais apresentaram maior possibilidade de caracterizar as diferenças de complexidade nas narrativas dos dois grupos de informantes. Para a obtenção dos resultados, calculou-se a média de cada métrica dos textos correspondentes a cada nível de escolaridade.

Resultados

As métricas de complexidade sintática se mostraram muito reveladoras no que diz respeito às diferenças de complexidade entre os textos produzidos pelos dois grupos de informantes. Os resultados de três dessas métricas são apresentados na tabela 1.

Tabela 1. Resultados de três métricas de complexidade sintática

	Ensino Fundamental	Ensino Superior
--	--------------------	-----------------

Distância na árvore de dependências	33,990112	52,351342
Proporção de orações subordinadas pela quantidade de orações do texto	0,231913	39,144816
Proporção de orações na voz passiva analítica em relação à quantidade de orações do texto	0,019933	0,031742

Como pode ser observado na tabela 1, as métricas das narrativas do corpus produzidas pelos alunos de ensino superior apresentam valores mais altos, indicando maior complexidade textual. Na métrica distância na árvore de dependências, existe uma relação entre a distância entre palavras e tempo de processamento. Quanto maior a distância, mais se exige da memória do destinatário do texto. Dessa forma, maior distância na árvore de dependências resulta em maior complexidade [Leal et al. 2021; Santucci et al. 2020]. No exemplo da figura 1, encontrado na narrativa ES9¹, pode-se observar uma grande distância entre a raiz da árvore (verbo “sair”) e a oração “parece ser alguma coisa assim”.

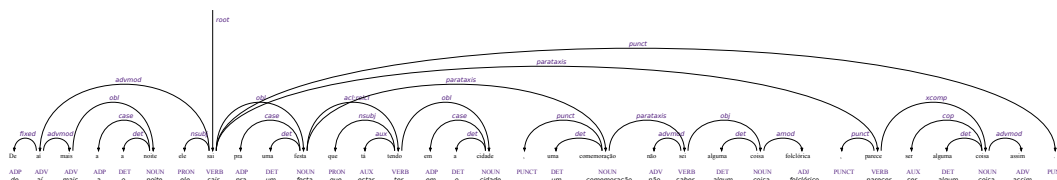


Figura 1. Distância na árvore de dependências

Pelo fato de as orações subordinadas serem estruturas mais complexas e que exigem maior esforço de processamento, a maior proporção desse tipo de construção indica maior complexidade textual [Leal et al. 2021; McNamara et al. 2014; Santucci et al. 2020]. No exemplo (1), retirado da narrativa ES1, encontram-se duas orações adjetivas (“que pode ser muito observada” e “que é Viva”) e uma oração completiva (“que ele acabou de presenciar”). No exemplo (2), retirado da narrativa ES3, encontram-se duas orações adverbiais (temporal: “quando já mais nem esperava encontrá-la”; causal: “porque nem mesmo sabia onde procurá-la”) e uma oração completiva (“onde procurá-la”).

(1) Uma antítese *que pode ser muito observada* seria o contraste entre a morte *que ele acabou de presenciar* e o nome do hotel, *que é Viva*.

(2) Então, *quando já mais nem esperava encontrá-la*, *porque nem mesmo sabia onde*

¹ ES: ensino superior. EF: ensino fundamental.

procurá-la, (...) encontrou a linda moça dos cachos dourados.

Uma maior proporção de orações na voz passiva analítica também indica maior complexidade textual, pois se trata de uma estrutura que as crianças adquirem mais tardiamente [Leal et al. 2021; McNamara et al. 2014]. No exemplo (3) a seguir, retirado da narrativa ES1, observam-se quatro construções passivas analíticas.

- (3) *Ele é impedido* de se aproximar dela por um senhor encarado e carrancudo,
e ela é retirada dele,
impedida de chegar perto dele por muitas pessoas
é trancafiada no quarto.

As métricas de diversidade lexical também apontam para uma maior complexidade das narrativas produzidas pelos alunos de curso superior. Na tabela 2, apresentam-se os resultados de duas dessas métricas.

Tabela 2. Resultados de duas métricas de diversidade lexical

	Ensino Fundamental	Ensino Superior
Proporção de <i>types</i> de substantivos em relação à quantidade de <i>tokens</i> de substantivos no texto	0,382287	0,56517
Proporção de <i>types</i> de verbos em relação à quantidade de <i>tokens</i> de verbos no texto	2,366978	250,618799

No caso das duas métricas, quanto maior a proporção de *types* em relação à quantidade de *tokens*, maior complexidade, ou seja, o produtor do texto demonstra conhecer uma maior diversidade de itens lexicais dessas duas classes de palavras. No caso dos substantivos, uma menor proporção de *types* indicaria menor complexidade textual pelo fato de a repetição de substantivos ser uma das formas mais simples de se construir uma cadeia de referência [Leal et al. 2021], como pode ser observado no exemplo (4), retirado da narrativa EF1, em que o produtor do texto, um aluno do ensino fundamental, repete os substantivos “homem”, “mulher” e “menina” para retomar os referentes anaforicamente.

- A história começa assim: é de *um homem* que chega numa cidade. (...) Chega lá daí de caravela. Chegando perto do hotel, vê um velório de *um homem*. Daí ele se
(4) apaixonou pela *mulher do homem que morreu*. Aí ele chega perto da *menina*, tinha uma festa lá. Aí chega perto da *menina*, o *pai da menina* chega junto com os segurança dele, daí ele não queria deixar ele ficar perto da *menina*.

Algumas métricas de informações morfosintáticas e de informações semânticas de palavras também demonstram maior complexidade nas narrativas dos informantes de

curso superior, como pode ser observado na tabela 2.

Tabela 3. Resultados de duas métricas de informações morfossintáticas e de duas métricas de informações semânticas de palavras

		Ensino Fundamental	Ensino Superior
Informações morfossintáticas	Proporção de pronomes relativos em relação à quantidade de pronomes do texto	0,243944	0,891063
	Proporção de verbos em relação à quantidade de palavras do texto	0,696373	1,439126
Informações semânticas	Proporção de nomes próprios em relação à quantidade de palavras do texto	0,021081	1,194962
	Proporção de substantivos abstratos em relação à quantidade de palavras do texto	0,196411	1,469733

Em se tratando das informações morfossintáticas, segundo Leal et al. (2021) e McNamara et al. (2014), a frequência mais alta de pronomes relativos indica maior complexidade pelo fato de esses pronomes introduzirem orações adjetivas, que elaboram o conteúdo de um sintagma nominal. No exemplo (5), encontrado na narrativa ES9, produzida por aluno de curso superior, o pronome relativo “que” retoma e elabora, na primeira ocorrência, o sintagma nominal “um marinheiro”; na segunda ocorrência, o sintagma nominal “uma fotografia no jornal”; na terceira ocorrência, o sintagma nominal “uma pessoa”.

A história é de um não sei parece ser um marinheiro *que* chega de navio numa
 (5) cidade e ele se hospeda num hotel, mas antes ele vê uma fotografia no jornal *que*
 parece ser de uma mulher, parece ser de uma pessoa *que* ele tá procurando.

Na métrica seguinte, a maior proporção de verbos plenos (não se consideram os verbos auxiliares na contagem) também indica maior complexidade pelo fato de os verbos constituírem orações.

No que diz respeito às informações semânticas, nomear entidades demanda mais

memória, motivo pelo qual uma maior proporção de substantivos próprios indica maior complexidade [Feng et al. 2010]. No exemplo (4), aqui retomado para facilitar a visualização, o produtor do texto, um aluno do ensino fundamental, utiliza substantivos comuns para nomear os personagens como “homem”, “mulher”, “menina”, “pai”. Já no exemplo (6), encontrado na narrativa ES10, o informante, um aluno de curso superior, utilizou nomes próprios “João” e “Maria” para designar os referentes.

- A história começa assim: é de *um homem* que chega numa cidade. (...) Chega lá daí de caravela. Chegando perto do hotel, vê um velório de *um homem*. Daí ele se
- (4) apaixonou pela *mulher do homem que morreu*. Aí ele chega perto da *menina*, tinha uma festa lá. Aí chega perto da *menina*, o *pai da menina* chega junto com os segurança dele, daí ele não queria deixar ele ficar perto da *menina*.

- João* voltou para sua cidade natal, num pequeno vilarejo nordestino. (...) Quando
- (6) decidiu voltar para o hotel, *João* encontrou uma moça muito bonita chamada *Maria* e por ela ele se apaixonou.

Em relação à outra métrica, o processamento de substantivos abstratos é mais trabalhoso do que o de substantivos concretos, motivo pelo qual uma maior proporção de substantivos abstratos indica maior complexidade [Leal et al. 2021].

Por fim, todos os índices de leitura demonstraram a maior complexidade das narrativas dos informantes de ensino superior. Ponomarenko e Evers (2022, p. 42) definem leitura como “potencial facilidade ou dificuldade de leitura de um texto”, levando em conta não apenas fatores linguísticos mas também o perfil do leitor pretendido pelo texto. De acordo com Yasseri, Kornai e Kertész (2012), a leitura é um dos principais temas relacionados à complexidade linguística. Apresentam-se, na tabela 4, os resultados do índice Gunning Fog [Gunning 1952] e do índice Flesch [Flesch 1979]. Esses dois índices são amplamente utilizados. Segundo Yasseri, Kornai e Kertész (2012), o índice Gunning Fox é uma das métricas mais confiáveis de leitura, e Branco et al. (2014b) afirmam que o índice Flesch é uma das métricas mais aceitas no que diz respeito à leitura.

Tabela 4. Resultados de dois índices de leitura

	Ensino Fundamental	Ensino Superior
Índice Gunning Fog	6,64537833333333	8,062712
Índice Flesch	71,5664383333333	60,57434

Segundo Štajner et al. (2012) e Leal et al. (2021), o índice Gunning Fog soma a quantidade média de palavras por sentença ao percentual de palavras com mais de duas sílabas (palavras difíceis) no texto e multiplica o resultado por 0,4. A fórmula é $0,4 \times (\text{comprimento médio das sentenças} + \text{palavras difíceis})$. Quanto maior o resultado, mais complexo o texto. Como se pode observar na tabela 4, as narrativas produzidas pelos alunos de ensino superior apresentam complexidade mais alta no que diz respeito ao

índice Gunning Fog.

Ainda segundo Štajner et al. (2012) e Leal et al. (2021), o índice Flesch relaciona o comprimento médio das sentenças e o número médio de sílabas por palavra. A fórmula é $248,835 - (1,015 \times \text{comprimento médio das sentenças}) - (84,6 \times \text{número médio de sílabas por palavra})$. Ao contrário do índice Gunning Fox, no índice Flesch, maior resultado da métrica indica menor complexidade textual. Dessa forma, como as narrativas produzidas pelos alunos de curso superior apresentam métrica mais baixa, elas são mais complexas do que as narrativas dos alunos de ensino fundamental de acordo com o índice Flesch.

Considerações finais

Neste trabalho, analisou-se um corpus formado por vinte narrativas orais (dez produzidas por estudantes do ensino fundamental e dez produzidas por estudantes do curso de Comunicação Social) com o objetivo de se verificar se a complexidade textual das narrativas aumenta conforme aumenta o nível de escolaridade. As onze métricas selecionadas demonstraram que as narrativas do corpus produzidas pelos alunos de curso superior apresentam maior complexidade textual do que as narrativas produzidas pelos alunos de ensino fundamental.

Nas métricas de complexidade sintática, foram encontradas, nas narrativas de curso superior, árvores de dependência com maior distância entre palavras relacionadas, uma proporção muito mais alta de orações subordinadas e uma maior proporção de construções na voz passiva analítica. Nas métricas de diversidade lexical, a proporção de *types* de substantivos e de verbos (em relação aos *tokens* dessas respectivas classes) também foi mais alta nas narrativas de curso superior.

Nas métricas de informação morfossintática das palavras, encontraram-se, nas narrativas dos alunos de ensino superior, uma maior proporção de pronomes relativos e uma maior proporção de verbos plenos. Em relação às métricas de informação semântica de palavras, foram encontradas, nas narrativas dos alunos de curso superior, uma maior proporção de nomes próprios bem como de substantivos abstratos.

Os índices de leitura, que levam em conta o comprimento das sentenças e a proporção de palavras difíceis, também indicaram maior complexidade nas narrativas dos alunos de curso superior.

Espera-se que este trabalho possa auxiliar no trabalho docente indicando algumas características que são esperadas de acordo com o grau de escolaridade. Conforme o aluno vai avançando nos níveis de ensino, presume-se que seus textos apresentem maior diversidade lexical (palavras diferentes e mais difíceis), maior complexidade sintática (uso de orações subordinadas e de pronomes relativos, orações mais longas, uso de construções passivas).

References

- Antonio, J. D. (1998). “Narrativas orais e narrativas escritas: a estrutura argumental preferida, e outras preferências”. In Faculdade de Ciências e Letras: Doutorado. Universidade Estadual Paulista Júlio de Mesquita Filho.
- Antonio, J. D. (2004). “Estrutura retórica e articulação de orações em narrativas orais e em narrativas escritas do português”. In Faculdade de Ciências e Letras: Doutorado. Universidade Estadual Paulista Júlio de Mesquita Filho.
- Evers, A. (2018). “A redação engaiolada: padrões lexicais e ensino de redação em cursos pré-vestibulares populares”. In Instituto de Letras: Doutorado. Universidade Federal do Rio Grande do Sul.
- Branco, A., Rodrigues, J., Costa, F., Silva, J. and Vaz, R. (2014a). Rolling out Text Categorization for Language Learning Assessment Supported by Language Technology. In *Computational Processing of the Portuguese Language: 11th International Conference, PROPOR 2014, Sao Carlos/SP, Brazil, October 6-8, 2014, Proceedings* (Vol. 8775, p. 256). Springer.
- Branco, A., Rodrigues, J., Costa, F., Silva, J. and Vaz, R. (2014b). Assessing automatic text classification for interactive language learning. In *International Conference on Information Society (i-Society 2014)* (pp. 70-78). IEEE.
- Goldman, S. R. and Lee, C. D. (2014). Text complexity: State of the art and the conundrums it raises. *The Elementary School Journal*, 115(2), 290-300.
- Feng, L., Jansche, M., Huenerfauth, M. and Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (p. 276-284).
- Flesch, Rudolf (1979). “How to write in plain English: A book for lawyers and consumers”, New York, Harper.
- Gunning, R. (1952). “The technique of clear writing”, McGraw-Hill, New York.
- Leal, S. E., Scarton, C. E., Cunha, A., Hartmann, N. S., Duran, M. S. and Aluísio, S. M. (2021) *NILC-Matrix Doc*. NILC-Matrix. Acesso em 19 mai 2023. Disponível em <<http://fw.nilc.icmc.usp.br:23380/matrixdoc>>.
- Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S. and Aluísio, S. M. (2022). NILC-Matrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *arXiv preprint arXiv:2201.03445*.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M. and Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Matrix, Cambridge, Cambridge University Press.

- Ponomarenko, G. L. and Evers, A. (2022). “Leiturabilidade e ensino: autores-base e seus trabalhos”, In *Acessibilidade textual e terminológica*, Edited by Maria José B. Finatto & Liana Braga Paraguassu, Uberlândia, Edufu, p. 41-71.
- Santucci, V., Santarelli, F., Forti, L. and Spina, S. (2020). Automatic classification of text complexity. *Applied Sciences*, 10(20), 7285.
- Sheehan, K. M., Flor, M. and Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility* (pp. 49-58).
- Štajner, S., Evans, R., Orasan, C. and Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility* (p. 14-22).
- Yasseri, T., Kornai, A. and Kertész, J. (2012). A practical approach to language complexity: a Wikipedia case study. *PloS one*, 7(11), e48386.

Aposições anafóricas e catafóricas no português e sua anotação no esquema *Universal Dependencies*

Magali Sanches Duran¹, Maria das Graças Volpe Nunes^{1,2}

¹Núcleo Interinstitucional de Linguística Computacional (NILC)

²Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)

magali.duran@uol.com.br, gracan@icmc.usp.br

Abstract. *This paper discusses the syntactic analysis of anaphoric and cataphoric appositions in Portuguese, which fall under the referentiality strategy called encapsulation. We present a review of studies that address the phenomenon and propose guidelines for annotating it from the perspective of dependency syntax, using labels from the Universal Dependencies approach.*

Resumo. *Este artigo discute a análise sintática de aposições anafóricas e catafóricas na língua portuguesa, as quais se enquadram na estratégia de referencialidade chamada de encapsulamento. Apresenta-se uma revisão de trabalhos que abordam o fenômeno e propõem-se diretrizes para anotá-lo sob a ótica da sintaxe de dependências, usando etiquetas da abordagem Universal Dependencies.*

1. Introdução

O exercício de anotação sintática de corpus apresenta ao anotador o desafio de reconhecer, na prática, fenômenos estudados pela linguística e descritos nas gramáticas. Porém, há alguns fenômenos que não são imediatamente reconhecidos, seja pelo fato de o anotador desconhecer sua descrição, seja pelo fato de tais fenômenos ainda não constarem dos manuais de anotação, ou, ainda, por não terem sido largamente descritos e consensualmente reconhecidos por gramáticos e linguistas.

Neste artigo discutem-se dois fenômenos que impuseram desafios ao projeto POeTiSA¹: as aposições catafóricas (exemplos 1, 2 e 3) e as aposições anafóricas (exemplos 4, 5 e 6), cujos correferentes estão indicados em negrito:

1. Com ares mediterrâneos, o Kez nasce de uma boa **ideia**: popularizar o bagel, aquele pão judeu redondo com um anel vazado no centro, denso, de crosta ligeiramente úmida.
2. Nas últimas semanas, ele aprendeu uma nova **técnica**: dedura o comentário para a empresa onde seu autor trabalha.
3. Há um **agravante**: como a JBS tem fábricas nos EUA, as propinas pagas no Brasil são uma violação da lei americana que proíbe empresas de lá de pagar suborno no exterior.
4. Até hoje não se sabe quem foi o autor - **o que** significa que ninguém foi punido.

¹<https://sites.google.com/icmc.usp.br/poetisa>

5. Depois, mesmo sofrendo o gol quando ainda faltavam mais de cinco minutos, soube segurar o resultado, **algo** que não vinha acontecendo em sua vida mais recente.
6. Passaram de R\$78,5 bilhões para R\$111,8 bilhões nos últimos dez anos, um **crescimento** real de 42% no período.

Esses fenômenos são denominados, em mais de uma gramática, “oração apositiva” (1, 2 e 3) e “aposto de oração” (4, 5 e 6). São casos, portanto, que orbitam a esfera do aposto, função na sintaxe que engloba muitos fenômenos e nem sempre é objeto de consenso entre os gramáticos e linguistas.

Todos os exemplos apresentados fazem parte de uma estratégia de referenciação chamada “encapsulamento”. No encapsulamento há uma palavra encapsuladora (pronome ou sintagma nominal) que é usada no lugar de toda uma oração por ela encapsulada. Segundo Koch (2008, p. 106) o encapsulamento é uma estratégia para construir novos objetos de discurso, que passam a constituir um referente para novas predicções. Quando ocorrem dentro de uma mesma sentença, tornam-se um problema de sintaxe. E é precisamente a análise sintática dessas construções de encapsulamento que inspirou as reflexões e propostas que se apresentam neste artigo. A análise utiliza a sintaxe de dependências (Tesnière, 2015) e o esquema de anotação da Universal Dependencies (UD) (De Marneffe *et al.* 2021; Nivre *et al.* 2020), exigindo basicamente duas definições em relação aos fenômenos em foco: qual etiqueta adotar para nomear a relação de dependência e quais são o *head* e o dependente da relação em cada caso.

Na Seção 2, apresentam-se brevemente a sintaxe de dependências e o esquema de anotação da abordagem UD, adotados como método de análise. Na Seção 3, revisa-se criticamente o tema das posições em gramáticas e estudos linguísticos. Na Seção 4 são discutidos exemplos de *cópus* e apresentadas propostas para sua anotação usando a sintaxe de dependências. Por fim, a Seção 5 traz conclusões e possibilidades de trabalhos futuros.

2. A sintaxe de dependências e a abordagem Universal Dependencies

A sintaxe de dependências faz uso de relações que ligam as palavras de uma sentença, duas a duas, sempre determinando qual é o *head* (ou governante) e qual é o dependente da relação. A UD é inspirada na sintaxe de dependências de Tesnière (1959 e 2015) e possui um conjunto de etiquetas para anotar a categoria morfossintática das palavras (17 *part-of-speech tags* ou *PoS tags*) e um conjunto de etiquetas para anotar as relações sintáticas entre palavras (37 *dependency relations* ou *deprel*).

A UD possui diretrizes² sobre como realizar a anotação e um fórum virtual onde são discutidas dúvidas e dificuldades de anotadores que adotaram a abordagem nas mais diversas línguas³. As diretrizes da UD já foram descritas e exemplificadas em manuais de anotação para língua portuguesa (Duran, 2021 e Duran, 2022) e há dois *cópus* de português brasileiro já disponíveis no site da UD: o Bosque-UD (Rademaker *et al.*

² Disponível em <https://universaldependencies.org/>

³ Até o momento da escrita deste artigo, 141 diferentes línguas apresentavam pelo menos um *cópus* anotado seguindo essa abordagem.

2017) e o PetroGold (Souza *et al.* 2021). Além disso, outras iniciativas de anotação de corpus de português nos moldes da UD encontram-se em andamento, como os relatados por Pardo *et al.* (2021), Oliveira *et al.* (2022) e Coneglian *et al.* (2022).

As etiquetas da UD guardam alguma equivalência com a nomenclatura gramatical brasileira, mas são muitas as diferenças entre os fenômenos cobertos e, por falta de espaço, não serão explicadas aqui. Para a discussão aqui instada, é essencial dizer que o arco das relações de dependência é direcional, partindo do *head* (ou governante) da relação e apontando sua flecha para o dependente da relação. Além disso, quando o head ou dependente de uma relação é uma oração, a ponta do arco correspondente à oração é colocada no núcleo do predicado; já quando o *head* ou dependente de uma relação é um sintagma, a ponta do arco correspondente é colocada no núcleo do sintagma. A Figura 1 ilustra uma árvore de dependências UD anotada com a ferramenta Arborator-Nilc (Miranda & Pardo, 2022).

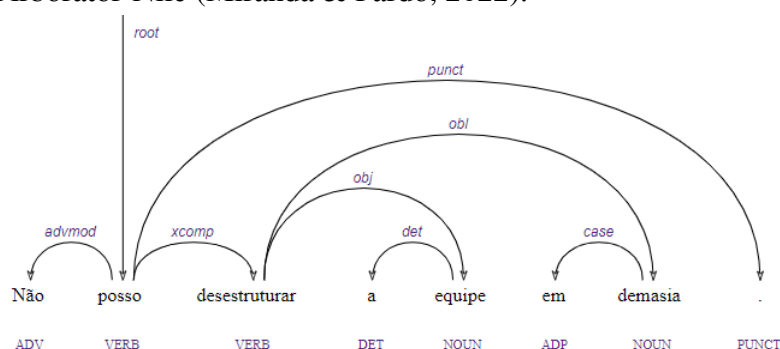


Fig. 1. Árvore de dependências de sentença no esquema UD

3. O estatuto sintático das aposições catafóricas e anafóricas

Nesta seção discute-se o tratamento das aposições catafóricas e anafóricas em gramáticas e estudos linguísticos do português, a fim de reunir subsídios para a anotação dessas construções dentro do esquema de anotação de corpus da UD.

3.1. Aposições catafóricas

Uma aposição catafórica é aquela em que um pronome ou substantivo genérico contido em uma oração é explicado ou comentado por uma outra oração. Esse fenômeno é comumente reconhecido nas gramáticas como uma oração que tem função de aposto (Cegalla, 2020, p. 385; Azeredo, 2013, p.139; Rocha Lima, 2011, p. 328; Faraco e Moura, 1994, p. 331). Essa oração é classificada como “subordinada substantiva apositiva”, antecedida por dois-pontos, e é a única das orações substantivas que não é selecionada pela semântica de verbos e nomes. Pode-se afirmar, portanto, que se trata de uma oração adjuntiva, ou seja, que não preenche uma lacuna da oração principal.

Teoricamente, as orações apositivas são orações dependentes e subordinam-se a um pronome ou substantivo contido na oração principal. Em sua forma finita, essas orações são introduzidas por uma conjunção integrante (“que” ou “se”) e, em sua forma nominal, são orações reduzidas (quase sempre de infinitivo). Azeredo (2013 p. 139) traz exemplos que se enquadram perfeitamente nessa definição, como o (7):

7. “Ele só pediu um favor: que o tirassem daquele hospital.”

Porém, alguns gramáticos chamam também de apositivas orações independentes, precedidas de dois-pontos, com verbo finito, ou seja, nem reduzidas e nem introduzidas por uma conjunção subordinativa. Isso pode ser observado nos exemplos 8 e 9, fornecidos por Cegalla (2020, p. 385), e nos exemplos 10, fornecido por Rocha Lima (2011, p. 332) e 11, fornecido por Faraco e Moura (1996, p.331).

8. "E confesso uma verdade: eu era um homem puro." (Povina Cavalcanti)
9. "A notícia veio de supetão: iam meter-me na escola." (Graciliano Ramos)
10. "Dei-lhe tudo: ofereci-lhe o meu nome; tornei-a dona de todo o meu dinheiro, elevei-a à minha posição social."
11. "Então aconteceu o pior: veio vento sul."

Nogueira e Leitão (2004) estudam as orações substantivas apositivas, "as quais estão associadas às estratégias de referenciação catafórica" e "cujo conteúdo é encapsulado e antecipado pela expressão referencial" (*op. cit.*, p.138).

A sentença 12 é fornecida pelas autoras como exemplo de oração apositiva, o que mostra que, para elas, assim como para Cegalla, Rocha Lima e Faraco e Moura (anteriormente citados), não importa se há ou não marcas explícitas de subordinação ligando a aposição à oração principal: em qualquer hipótese a aposição catafórica em forma oracional será classificada como oração apositiva.

12. "Só vejo uma situação: o Governo do PT elegeu-se pregando uma coisa e atualmente faz totalmente o inverso." (Nogueira e Leitão, 2004, p. 139)

Na sentença 12, "situação" é a palavra encapsuladora e "o Governo do PT ... inverso" é a oração coordenada encapsulada, conforme ilustrado na Figura 2.

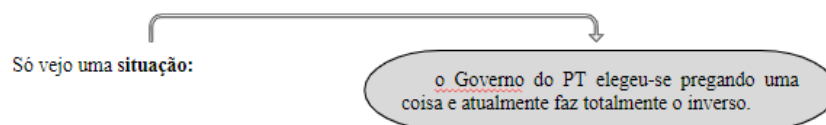


Fig. 2. Encapsulamento catafórico (da esquerda para a direita)

3.2. Aposições anafóricas

A relação inversa também pode ocorrer: uma aposição que contém um pronome ou substantivo que comenta o conteúdo da oração principal. Alguns gramáticos, como Bechara (1999), denominam esse tipo de aposição de "aposto de oração", mas há gramáticos que usam outra denominação, como Nougé (2015), que o chama de "aposto resumitivo", e ainda gramáticos, como Rocha Lima (2011), que reconhecem a possibilidade de um aposto que se refira a uma oração, porém não o nomeiam.

A ideia de que a "oração apositiva" é o inverso do "aposto de oração" parece encontrar respaldo nos estudos de linguística textual e discursiva acerca das estratégias de encapsulamento usadas para construir coesão textual. Francis (1994, p. 98) faz a distinção entre o "*advance label*", que funciona cataforicamente, e "*retrospective label*", que funciona anaforicamente. Observa-se que o "*advance label*" produz o encapsulamento catafórico (para frente), o que engloba o conceito estendido de "oração apositiva", ao passo que o "*retrospective label*" produz o "encapsulamento anafórico" (para trás), o que engloba o conceito de "aposto de oração".

As construções chamadas de “aposto de oração” são tema dos estudos de Sousa (2016). Ao descrever o aposto de oração, Sousa afirma que a função pode ser desempenhada por um substantivo, seguido de oração adjetiva, ou pelo pronome “o”, seguido de uma oração adjetiva, ou simplesmente por um substantivo. As sentenças 13, 14 e 15, fornecidas no trabalho da autora, exemplificam essas três realizações:

13. “Paradoxalmente, a partir da década de 90, a prática de extorsão mediante sequestro foi sensivelmente maior do que antes, fato que se repetiu, aliás, com outros tantos delitos mais ou menos hediondos.”
14. “(...) atenuam responsabilidades por meio do argumento de que também os outros partidos cometeram pecados e que destes não se falou o quanto era preciso, argumento insuficiente (mesmo se a imputação é legítima), já que, além do tamanho da operação, há, no caso do PT, uma circunstância agravante pelo fato de se tratar de um partido que se apresenta como modelo de virtude cívica.”
15. “Mais do que planilhas e números, estamos falando de pessoas, na ativa ou não, que terão maior poder de compra, o que gera, de imediato, melhoria na qualidade de vida do cidadão.”

No exemplo 13 há claramente uma oração encapsulada e um substantivo encapsulador (“fato”), como mostra a Figura 3.

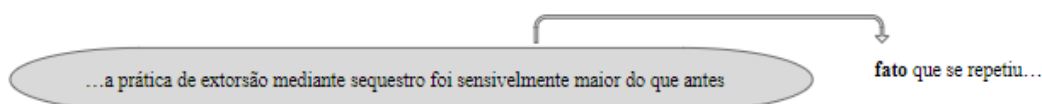


Fig. 3. Encapsulamento anafórico (da direita para a esquerda)

Já o exemplo 14, ao que parece, não constitui um aposto de oração, mas um aposto comum, que liga “argumento de que...” e “argumento insuficiente”.

E quanto ao exemplo 15, não parece ter comportamento similar ao do exemplo 13. Enquanto um substantivo ou pronome indefinido pode ser adjetivado (exemplo 16), o pronome “o” não o pode (exemplo 17).

16. (...) a prática de extorsão mediante sequestro foi sensivelmente maior do que antes, **fato/algo interessante** que se repetiu (...)
17. *(...) terão maior poder de compra, **o interessante que** gera (...)

Além disso, o pronome relativo “que” da oração adjetiva restritiva cujo antecedente é um substantivo ou pronome indefinido pode ser preposicionado (exemplo 18), enquanto o pronome “que” que acompanha o pronome “o” não pode (19).

18. (...) a prática de extorsão mediante sequestro foi sensivelmente maior do que antes, **fato/algo de que** nem todos se lembram.
19. *(...) terão maior poder de compra, **o de que** todos se orgulham.

Na verdade, se houvesse uma preposição, ela antecederia “o que”, como mostra o exemplo 20.

20. (...) terão maior poder de compra, do que todos se orgulham.

Ao que parece, os pronomes “o” e “que” nesse uso são indissociáveis, formando uma locução pronominal, análoga ao pronome relativo “o qual” e suas flexões. Mas, se

“o que” for um pronome relativo, como se chamaria uma oração relativa que acrescenta informação a uma outra oração? Pelo fato de modificar um predicado e de ser adjuntiva, poderia se chamar “oração subordinada adverbial relativa”? Independentemente da denominação, o que distingue o aposto de oração da oração relativa é o fato de que o dependente da relação de dependência passa a ser o predicado da oração relativa, como mostrado na Figura 4.

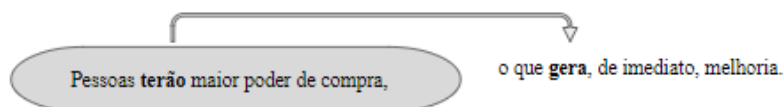


Fig. 4. Encapsulador e encapsulado em forma de oração

É interessante observar ainda que, se a anotação não fosse de relações sintáticas, mas de relações de correferência, a relação sempre partiria do encapsulador em direção ao encapsulado, independentemente de o encapsulador estar à direita ou à esquerda do encapsulado.

4. Relações de dependência entre aposições catafóricas e anafóricas

Inicialmente justifica-se a utilização do termo “aposição” e não “aposto” para designar as construções em foco pelo fato de que, no esquema de anotação UD, a relação de dependência **appos** só pode ser atribuída a relações entre substantivos (próprios e/ou comuns) e entre substantivos e pronomes, e apenas na direção da esquerda para a direita. Como o esquema UD tem como objetivo servir a várias línguas, não existe flexibilidade para utilizar a relação **appos** para anotar o que na literatura é descrito como “oração apositiva” e “aposto de oração”. De fato, isso geraria um problema de inconsistência entre as línguas anotadas no esquema, prejudicando a comparação de suas estruturas sintáticas.

Pelos mesmos motivos, também não é permitido criar novas etiquetas de relações de dependência, a não ser de forma consensual entre todos os cientistas que participam da iniciativa. Quando ocorrem acordos de mudanças no conjunto de etiquetas ou na forma de empregá-las, as diretrizes da UD são alteradas e todos os inscritos recebem notificação sobre as alterações. Como é alto o custo de reanotar os 245⁴ corpúscos já disponíveis na UD (em 141 línguas), mudanças no esquema são raras. Isso não impede que haja muita discussão entre aqueles que adotam o esquema, e a questão do aposto em que um dos termos é uma oração é um tema recorrente⁵.

Diante das restrições do esquema de anotação, acredita-se que a relação de dependência mais adequada para anotar as aposições catafóricas e anafóricas na UD seja

⁴ Dado de 15/05/2023

⁵ As discussões acerca do aposto oracional podem ser encontradas nos seguintes links:

<https://github.com/UniversalDependencies/docs/issues/762>

<https://github.com/UniversalDependencies/docs/issues/751>

<https://github.com/UniversalDependencies/docs/issues/523>

a **parataxis**. Essa opção parece ser bem adequada quando não há marcas de subordinação entre o termo encapsulador e o termo encapsulado nas aposições, porém é uma solução pouco defensável sob o ponto de vista das gramáticas, já que a parataxe não deveria se confundir com a hipotaxe (subordinação). Mas é uma concessão que se faz enquanto não há melhor opção.

Assim, a aposição catafórica anotada com a relação **parataxis** teria como *head* o nominal que encapsula uma oração, como ilustrado na Figura 5, que mostra parte da árvore sintática do exemplo (1).

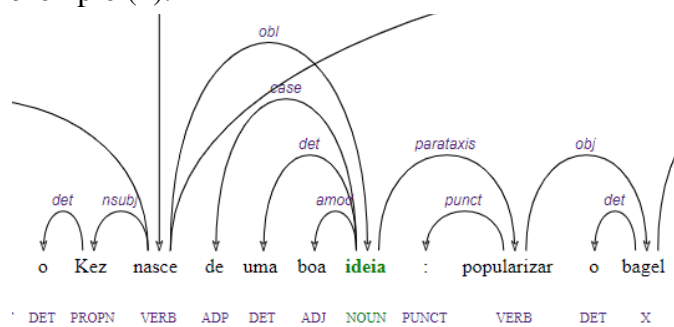


Fig. 5. Anotação no esquema UD de aposição catafórica (oração apositiva)

Já a aposição anafórica anotada com a relação **parataxis** tem como *head* o predicado da oração encapsulada e, como dependente, a palavra encapsuladora, como mostra a Figura 6 (parte da árvore sintática do exemplo 5).

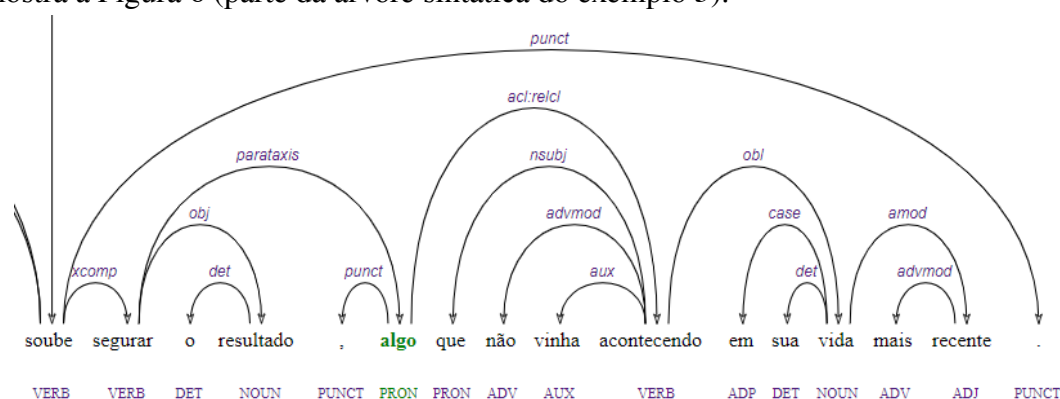


Fig. 6. Anotação no esquema UD de aposição anafórica (aposto de oração)

Contudo, no caso das aposições anafóricas iniciadas por “o que”, decidiu-se usar uma outra anotação, reconhecendo os dois pronomes como uma locução que funciona como um pronome relativo. Para isso, os dois pronomes são unidos pela relação de dependência **fixed**, utilizada para palavras funcionais constituídas de mais de um token. Nesse caso, a relação de dependência não se dá entre a oração encapsulada e o pronome encapsulador, mas entre a oração encapsulada e a oração que contém o encapsulador, à semelhança do que ocorre nas orações relativas cujo antecedente é um nominal.

Inicialmente, optou-se por utilizar a relação **parataxis** inclusive nesse caso de relativa, pois a relativa **acl:relcl**, prevista na UD, só se aplica a antecedentes nominais. Porém, tem se mostrado ser plausível a possibilidade de reconhecer, como oração adverbial relativa, as aposições iniciadas por “o que” cujo antecedente é oracional. Isso cumpriria dois requisitos: reconhecer seu estatuto de oração subordinada e seu estatuto

de adjunção. A Figura 7 ilustra a anotação da aposição anafórica encabeçada por “o que” na condição de pronome relativo.

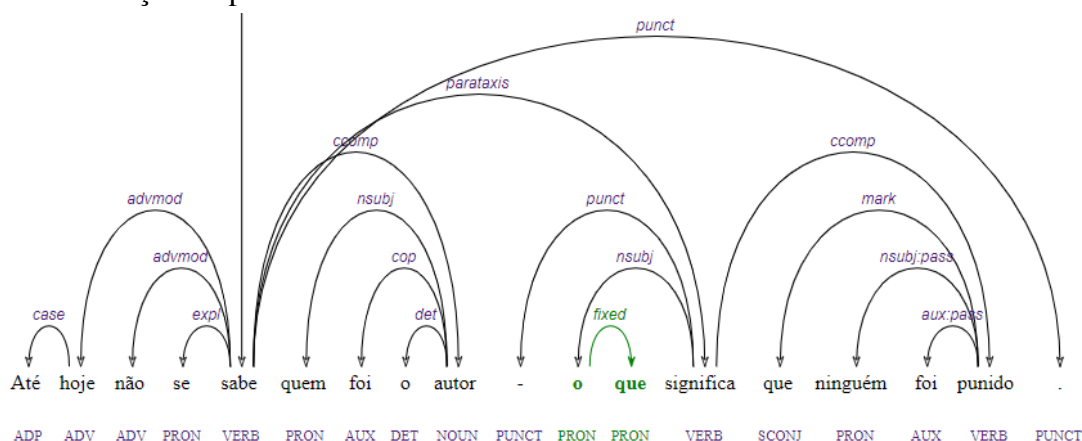


Fig. 7. Anotação no esquema UD de oração relativa com antecedente oracional

Cabe esclarecer que a UD permite que sejam acrescentadas sub-relações a fim de discriminar tipos de uma mesma relação de dependência. Assim, caso seja interessante para o projeto de anotação, a relação **parataxis** utilizada para anotar as aposições catafóricas e anafóricas poderia receber uma sub-relação, como **parataxis:appos**, por exemplo. O mesmo poderia ser feito no caso das relativas encabeçadas por “o que”, que poderiam ser anotadas como **parataxis:relcl** (*relcl* de *relative clause*) ou como **advcl:relcl** (adverbial relativa).

5. Conclusões e trabalhos futuros

As construções apositivas catafóricas e anafóricas, denominadas, respectivamente, orações apositivas e apostos de oração em gramáticas e estudos linguísticos, motivaram uma série de reflexões neste artigo. Questionou-se o fato de orações sem marcas de subordinação estarem sendo reconhecidas como orações subordinadas apositivas, o que sugere que se trata de um tema que merece estudos mais profundos sob o ponto de vista sintático. Da mesma forma, questionou-se a anotação de orações encabeçadas por “o que” que encapsulam o conteúdo de uma oração precedente, pois testes demonstraram que “o que” se comporta como uma locução e não como um encapsulador nominal comum (“o”) seguido de oração relativa introduzida por “que”. Propõe-se reconhecer “o que” como uma expressão fixa com função de pronome relativo, à semelhança de “o qual” e suas flexões, quando encabeçam orações relativas.

Um trabalho futuro é discutir com outros grupos que anotam corpus seguindo o esquema da UD a possibilidade de reconhecer as orações relativas encabeçadas por “o que”, e cujo antecedente é uma oração, como orações adverbiais relativas.

Uma vez que se tenha chegado a um consenso, com outros grupos que empregam o esquema UD em português, a respeito das relações a serem adotadas na anotação das aposições catafóricas e anafóricas, as decisões deverão ser incluídas nos manuais que contêm as diretrizes de anotação, a fim de promover a disseminação das respectivas análises entre os anotadores.

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

- Azeredo, J. C. S. Fundamentos de Gramática do Português. E-book. Jorge Zahar Editores, Rio de Janeiro, 2013.
- Bechara, E. Moderna Gramática Portuguesa. 39ª edição. Editora Nova Fronteira: Rio de Janeiro, 2019.
- Castilho, A. T. Gramática do Português Brasileiro. Editora Contexto, São Paulo, 2010.
- Cegalla, D. P. Novíssima Gramática da Língua Portuguesa. 49ª edição. Companhia Editora Nacional, 2020.
- Coneglian, A. V. L.; Guimarães, A. L. A. R.; Ferreira, T. C.; Pagano, A. S. Anotação de textos não canônicos: um estudo exploratório de Grande sertão: veredas pelas dependências universais. *In: Proceedings of the Universal Dependencies Brazilian Festival*, pp. 1-11. Association for Computational Linguistics, Fortaleza, Brazil, 2022.
- Cunha, C. F.; Lindley Cintra, L. F. Nova gramática do Português contemporâneo. 7a edição. Lexikon Editora Digital, Rio de Janeiro, 2017.
- Duran, M. S. 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, Universidade de São Paulo, 55p., Set. 2021. Disponível em: https://drive.google.com/file/d/1BddPswN-_Ioo-A5GsldA1cO1kqbcCahb/view?usp=sharing
- Duran, M. S. Manual de Anotação de Relações de Dependência – Versão Revisada e Estendida. Relatório Técnico do ICMC 440. ICMC-USP. São Carlos-SP, Out 2022, 166p. Disponível em: <https://drive.google.com/file/d/1ile8Wfxu1qdrZOmLGqkvVuQ4fXvHgVMo/view?usp=sharing>
- Faraco, C. E.; Moura, F. M.. Gramática. Editora Ática, São Paulo, 1994.
- Francis, G. Labelling discourse: an aspect of nominal-group lexical cohesion. In: Coulthard, M. (Ed.), *Advances in Written Text Analysis*. Routledge, London, p. 83-101, 1994.

- Marneffe, M.; Manning, C.; Nivre, J.; Zeman, D. Universal Dependencies. *Computational Linguistics* 47 (2) p. 255-308. MIT PRESS, 2021.
- Miranda, L. G. M.; Pardo, T. A. S. An Improved and Extended Annotation Tool for Universal Dependencies-based Treebank Construction. *In: Proceedings of the PROPOR Demonstrations Workshop*, p. 1-3, 2022.
- Neves, M. H. de M. Gramática de Usos do Português. Ed. Unesp, São Paulo, 2000.
- Nivre, J.; Marneffe, M.; Ginter, F.; Hajič, J.; Manning, C.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *In: Proceedings of the 12nd International Conference on Language Resources and Evaluation (LREC 2020)*, p. 4034-4043, 2020.
- Nogueira, M. T.; Leitão, R. J. A oração substantiva apositiva: aspectos textual-discursivos. *Veredas*, v.8, n.1 e n.2, p. 137-151. Juiz de Fora, 2004.
- Nougué, C. Suma Gramatical da Língua Portuguesa: Gramática Geral e Avançada. É Realizações, São Paulo, 2015.
- Oliveira, L. F. A.; Pagano, A.; Oliveira, L. E. S.; Moro, C. Challenges in Annotating a Treebank of Clinical Narratives in Brazilian Portuguese. *In: Computational Processing of the Portuguese Language*, Volume 13208, 2022.
- Pardo, T. A. S.; Duran, M. S.; Lopes, L.; Di Felippo, A.; Roman, N. T.; Nunes, M. G. V. Portinari - A large multi-genre treebank for Brazilian Portuguese. *In: Proceedings of the XIV Symposium in Information and Human Language (STIL 2021)*, p. 1-10, 2021.
- Perini, M.. Gramática Descritiva do Português. 2ª ed., 380 p. Ática, São Paulo, 1996.
- Rademaker, A.; Chalub, F.; Real, L.; Freitas, C.; Bick, E.; de Paiva, V. (2017). Universal dependencies for Portuguese. *In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, p. 197–206, 2017.
- Rocha-Lima, C. H. Gramática normativa da língua portuguesa. Editora José Olympio, São Paulo, 2010.
- Sousa, R. S. N A aposição encapsuladora em artigos de opinião no português. Tese (Doutorado) – Universidade Federal do Ceará, Programa de Pós-Graduação em Linguística, Fortaleza (CE), 2016.
- Souza, E; Silveira, A.; Cavalcanti, T.; Castro, M. C.; Freitas, C.. PetroGold – Corpus padrão ouro para o domínio do petróleo. *In: Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 13. p. 29-38. Sociedade Brasileira de Computação, Porto Alegre, 2021.
- Tesnière, L. (2015). Elements of Structural Syntax. Tradução de Osborne, Timothy; Kahane, Sylvain. John Benjamins, Amsterdam, 2015.

Um pronome com muitas funções: Descrição e resultados da anotação do pronome *–se* em um *treebank* segundo o esquema *Universal Dependencies* (UD) para Português

Elvis de Souza¹, Cláudia Freitas²

¹Departamento de Letras – PUC-Rio
Lab. Inteligência Computacional Aplicada – PUC-Rio

²Departamento de Letras – PUC-Rio

elvis.desouza99@gmail.com, claudiafreitas@puc-rio.br

Abstract. *In this work, we conducted a linguistic description and reported the annotation process of the pronoun –se in the PetroGold v3 treebank [de Souza 2023]. Special attention to the pronoun –se is justified by the need to correctly annotate cases where the pronoun indicates subject indeterminacy, synthetic passive voice, or pronominal verb, recognizing its relevance for various NLP tasks. As a result, we categorized the 1,960 occurrences of “se” in the corpus according to their syntactic class and presented the verbs associated with each type (or more than one type) of the pronoun –se.*

Resumo. *Neste trabalho, realizamos uma descrição linguística e relatamos o processo de anotação do pronome –se no treebank PetroGold v3 [de Souza 2023]. A atenção especial ao pronome –se se justifica pela necessidade de anotar corretamente os casos em que o pronome indica indeterminação do sujeito, voz passiva sintética ou verbo pronominal, reconhecendo sua importância para diversas tarefas de PLN. Como resultados, discriminamos as 1.960 ocorrências do “se” no corpus por classe sintática e apresentamos os verbos que se associam a cada um (ou mais de um) dos tipos do pronome –se.*

1. Introdução

Neste trabalho, realizamos uma descrição linguística e relatamos o processo de anotação do pronome *–se* no *treebank* PetroGold v3 [de Souza 2023] (250.605 *tokens*). Apresentamos a metodologia – ferramentas e procedimentos – empregada na anotação e na sua avaliação, as decisões linguísticas subjacentes à anotação e discutimos os resultados sob diferentes pontos de vista.

O *treebank* integra o projeto *Universal Dependencies* (UD) [de Marneffe et al. 2021], e portanto seguimos as diretrizes de UD para conduzir a anotação. Enquanto as diretrizes, por um lado, permitem uma anotação muito simples para o pronome *–se*, agrupando todos sob uma mesma etiqueta sintática, por outro lado há também a possibilidade, igualmente prevista nas diretrizes, de utilizar etiquetas mais específicas para classificar o pronome. Utilizamos gramáticas do português [Bagnó 2012, Bechara 2012, Cunha and Cintra 2016] para ajudar a realizar a anotação dos tipos específicos de pronome *–se* de uma maneira que fosse respaldada tanto pelas diretrizes do projeto UD quanto pela tradição gramatical da língua portuguesa.

A atenção especial ao pronome *–se* se justifica, em termos gerais, pela necessidade de garantir a confiança na anotação padrão ouro do *corpus* e pelo desejo de alinhar as diretivas do projeto UD às gramáticas do português, facilitando o diálogo entre estudos linguístico-descritivos para a língua portuguesa e tarefas linguístico-computacionais. Em termos específicos, há também a necessidade de anotar corretamente os casos em que o pronome *–se* indica indeterminação do sujeito, voz passiva sintética ou verbo pronominal, reconhecendo que é fundamental para a identificação dos argumentos de um verbo, sendo importante para as tarefas de identificação de papéis semânticos e extração de informação, que são dificultadas em frases de sujeito omitido (pelo uso do *–se*, por exemplo), um problema já abordado anteriormente [Duran and Aluísio 2011, Hartmann et al. 2014, Freitas and de Souza 2021].

A identificação de sujeitos ocultos, por exemplo, só pode ser realizada satisfatoriamente quando o pronome *–se* é corretamente identificado porque, em frases de sujeito indeterminado, não há sujeito a ser encontrado; quando o pronome indica voz passiva sintética ou verbo pronominal, porém, sabe-se que a ausência de um sujeito sintático é resultado de elipse (e que este não é o sujeito semântico nas orações de voz passiva), e que deve poder ser encontrado no texto.

Além de discutir as dificuldades e os resultados da anotação, apresentamos também o impacto das revisões do pronome *–se* na geração de um modelo de aprendizado de máquina. Por fim, disponibilizamos três recursos lexicais processáveis computacionalmente, que podem auxiliar em outros projetos de anotação e em tarefas de processamento de linguagem natural (PLN).

2. Metodologia

A anotação do pronome *–se* foi realizada no *corpus* PetroGold, um *treebank* padrão ouro (250.605 *tokens*) composto por 19 teses e dissertações do domínio do petróleo. A revisão da anotação do pronome foi publicada na terceira versão do recurso, disponível tanto nas páginas do projeto Petrolês quanto do projeto *Universal Dependencies*¹.

O *corpus* integra o projeto *Universal Dependencies* (UD) [de Marneffe et al. 2021], uma iniciativa para padronização da anotação morfosintática e para disponibilização de *datasets* e ferramentas para diversas línguas. Como consequência, o mais importante no processo de anotação do PetroGold é garantir que as diretivas de anotação empregadas na tarefa se alinham às diretivas do projeto.

Antes da revisão relatada neste trabalho, o *corpus* continha uma única etiqueta de função sintática para designar todos os pronomes *–se*: a etiqueta *expl*, para pronomes expletivos. Essa anotação tinha origem no modelo que primeiro anotou o recurso, treinado majoritariamente no *corpus* Bosque-UD [Rademaker et al. 2017], um *treebank* composto por textos jornalísticos que também integra o projeto UD. A anotação não é incorreta segundo as diretivas do projeto, porém há a possibilidade – prevista nas diretivas – de utilizar etiquetas específicas para diferenciar o tipo de pronome sendo empregado em cada uma das ocorrências no *corpus*. Realizamos essa especificação com a ajuda de gramáticas do português (doravante GT, ou gramáticas tradicionais), uma vez que as diretivas do

¹Os endereços dos projetos são, respectivamente, <https://petroles.puc-rio.ai> e <https://universaldependencies.org>. Acesso em 15 de jun. 2023.

projeto UD ainda não têm um detalhamento da descrição para casos específicos da língua portuguesa.

Dessa forma, o processo de anotação do pronome *–se* no PetroGold consistiu em (1) revisar a anotação do pronome *–se*, garantindo que todas as ocorrências anotadas como pronome (e com a etiqueta de função *expl*) são de fato os pronomes que queremos anotar, e (2) especificar o tipo de pronome *–se*, diferenciando quando (a) o pronome tem função de indeterminação do sujeito, (b) é utilizado para empregar voz passiva sintética ou (c) indica um verbo sendo usado pronominalmente (cada um dos três usos corresponde a etiquetas de função sintática diferentes para o pronome).

A etapa (1), relativa à revisão da anotação da classe gramatical, função sintática e lema do “se” para garantir que eram de fato pronomes – e não conjunções subordinativas, por exemplo –, foi realizada utilizando a ferramenta [de Souza and Freitas 2021], um ambiente de busca e revisão de *corpora* anotados morfossintaticamente. Realizamos uma busca por todas as formas da palavra “se” e, para facilitar a revisão de todas as ocorrências, pedimos na ferramenta a distribuição dos casos de acordo com a anotação original de classe gramatical, função sintática e lema, tornando menos trabalhosa a identificação humana de erros uma vez que as ocorrências haviam sido agrupadas nas buscas. Identificados os casos errados, foram corrigidos em lote por meio de regras de correção, uma das possibilidades da ferramenta.

A etapa (2), relativa à especificação do tipo de pronome *–se* (índice de indeterminação, pronome apassivador ou partícula integrante do verbo), foi realizada de forma semelhante à etapa (1), agrupando frases semelhantes – dessa vez, tentamos agrupar os casos de indeterminação de sujeito, verbo pronominal e voz passiva sintética pelas características do verbo ao qual o pronome *–se* se associa. Pedimos à ferramenta pela distribuição dos lemas verbais dos quais o *token* “se” é dependente sintaticamente, o que nos permitiu visualizar a lista de verbos, agrupá-los quanto à transitividade, e então utilizamos o agrupamento para facilitar a leitura e a análise de todas as frases para verificar qual seria a anotação correta para o pronome nos contextos respectivos².

Por fim, para verificar o impacto das revisões e da especificação das funções do pronome *–se* no aprendizado automático – ou, na generalização de casos –, comparamos a saída de um modelo treinado no *dataset* antes das revisões e de outro modelo treinado no *dataset* após as revisões. A hipótese em que se baseia essa metodologia é a de que a qualidade do modelo gerado indica a possibilidade de generalização das análises linguísticas e, indiretamente, o nível de consistência dessas mesmas análises. Mas é importante notar que baixa qualidade do modelo (desempenho fraco) não significa, necessariamente, inconsistência de análise, uma vez que análises podem ser consistentes e de difícil generalização. Além disso, no caso da anotação do pronome *–se*, a análise dos resultados do modelo indica também as dificuldades que a implementação de três novas etiquetas impõem ao modelo de anotação automática

A ferramenta utilizada para treinar o modelo foi o UDPipe [Straka et al. 2016], e as métricas empregadas na avaliação da qualidade do modelo foram as métricas da avaliação conjunta do CoNLL de 2018 [Zeman et al. 2018]. Daremos ênfase aos resulta-

²Para reproduzir a metodologia de anotação, quem lê pode se referir a [de Souza 2023], onde é apresentado o passo a passo utilizado para especificar o pronome *–se*.

dos de UPOS (*Universal Part-Of-Speech Score*, que mede os acertos de classe gramatical), LEMMA (que mede os acertos de lematização) e LAS (*Labeled Attachment Score*, que mede os acertos de encaixe das dependências sintáticas e de tipo de relação sintática).

3. Classes do *–se*

Para especificar o pronome *–se* foram utilizados três subtipos da relação *expl* – *expl:impers*, *expl:pass* e *expl:pv* –, além das classes para objetos – *obj* e *iobj*³. Em todos esses casos, o *token* “se” é o que recebe a etiqueta de relação sintática específica, e a sua anotação de classe gramatical é “PRON”.

Embora o termo “impers” da classe “*expl:impers*” diga respeito à impessoalização (*impersonal*, em inglês), entendemos, junto com [Bagno 2012], que tanto a indeterminação do sujeito quanto a voz passiva sintética são estratégias utilizadas para impessoalizar a oração. No caso da indeterminação do sujeito, a impessoalização ocorre pela supressão do sujeito, que não pode ser recuperado na oração, ao passo que no caso da voz passiva sintética a impessoalização ocorre pelo deslocamento do objeto direto para a posição de sujeito paciente, sendo que o agente também não é recuperável dentro da oração. Estudos em diferentes abordagens teóricas também defendem a função de indeterminação do *–se* nos casos que gramáticas normativas classificam como partícula apassivadora [Vieira and de Sá 2015, Lopes and Namiuti-Temponi 2017, dos Santos Silva 2021]. No entanto, uma vez que o conjunto de etiquetas UD prevê a tripla diferenciação, optamos por mantê-la. Outro argumento para a separação “tradicional” entre passiva e indeterminação é justamente viabilizar estudos descritivos como os citados. Por fim, e de um ponto de vista prático, se o interesse estiver apenas na distinção entre indeterminação e demais usos – relevante, por exemplo, na anotação de papéis semânticos – é possível dar um tratamento unificado a *expl:impers* e *expl:pass*.

Assim, estamos utilizando a etiqueta *expl:impers* especificamente para os casos de indeterminação de sujeito e a etiqueta *expl:pass* para os casos de voz passiva sintética. A diferenciação entre ambas as classes, portanto, não se dá pela noção de “impessoalização”, mas por critérios gramaticais que serão explicados a seguir⁴.

***expl:impers* – índice de indeterminação do sujeito:** Comumente, o verbo cujo sujeito está indeterminado é intransitivo ou transitivo indireto. Um dos requisitos para a anotação de indeterminação do sujeito é a ausência de um sujeito sintático para o verbo a que o “se” está associado.

1. Considerando-se que o ciclo de o motor é realizado a cada duas rotações completas de o eixo de manivelas, **chega-se** a expressão de a massa de combustível (em kg) utilizada em cada ciclo: (3.1)

³A palavra “se” pode ainda ser uma conjunção subordinativa ou um nome próprio – abreviação de sudeste ou Sergipe (SE). Como não são casos de pronome *–se*, não serão tratados neste trabalho.

⁴As diretivas apresentadas nesta seção resumem, devido a limitações de espaço, os critérios utilizados para anotar o pronome *–se* no *corpus* PetroGold. Para um detalhamento maior sobre casos específicos e dúvidas que surgiram durante a anotação, quem lê pode se referir ao trabalho [de Souza 2023], onde as diretivas empregadas são exploradas mais extensamente.

expl:pass – pronome apassivador: É condição para a ocorrência de voz passiva sintética um verbo com transitividade direta ou transitividade direta e indireta segundo a GT. Em outras palavras, o requisito é a presença de um objeto direto, o qual, na transformação para voz passiva sintética, será anotado como sujeito do ponto de vista sintático, embora seja paciente do ponto de vista semântico.

Diferentemente da oração em que há indeterminação do sujeito, nesta há um sujeito sintático, anotado como *nsubj:pass* – sujeito paciente.

2. Para a análise de as argilas estudadas, **seguiu-se** o seguinte procedimento; **secou-se** as argilas em 39 uma estufa a 80°C durante 18 horas, e após terem sido retiradas de a estufa, foram moídas em um moinho de bolas durante 18 horas.

Na frase 2, o fato de que “argilas”, sendo sujeito da oração, não concorda em número com “secou”, não é um bom critério para definir se houve indeterminação do sujeito ou voz passiva sintética. Entendemos, assim como [Bagno 2012], que em ambos os casos o objetivo é impessoalizar a oração, de tal maneira que já se tornou usual conjugar o verbo na terceira pessoa do singular, mesmo que o fenômeno empregado seja o da voz passiva sintética. Assim, embora segundo a GT tenha ocorrido um erro de concordância verbal, ele é explicado pela intenção do autor, que não distingue entre um sujeito paciente e uma oração de sujeito indeterminado na hora de impessoalizá-la, e o *–se* foi anotado como *expl:pass*.

expl:pv – verbo pronominal: Indica que o *–se* está associado a um verbo pronominal e recebe o nome, na GT, de partícula integrante do verbo. Essa é uma análise que, para nós, assim como as demais categorias, depende da utilização do verbo em cada frase, não sendo característica intrínseca dos verbos exigir ou não o pronome “se” (veja-se o caso dos verbos “pronominais acidentais”, por exemplo). O verbo está sendo usado de forma pronominal quando há um sujeito sintático e ele não é paciente da ação verbal.

3. Dentre os parâmetros reológicos mais usuais, destaca se a viscosidade, que **se refere** a a resistência que uma substância apresenta a o fluxo (...)

Em alguns casos, o sujeito não é nem paciente e nem agente, quando um verbo causativo tornou-se incoativo pelo uso do pronome *–se*, como na frase “O esporte popularizou-se”, levantada em [Duran et al. 2013]. Nela, o verbo, originalmente transitivo direto, está sendo empregado no aspecto incoativo – “o esporte ficou popular” – indicando uma mudança de estado e, por isso, “esporte” não é nem agente nem paciente de um verbo de ação, mas “sede” da mudança de estado indicada pelo verbo (de estado), termo empregado por [Cunha and Cintra 2016]. Frases do tipo foram anotadas como pronominais⁵.

obj / iobj – objeto direto ou indireto: A palavra “se” pode ainda ser anotada como objeto direto ou indireto quando a ação do verbo se estende à terceira pessoa do singular ou plural, na forma do pronome “se”. Nesses casos, o sujeito sintático também é agente e paciente da ação ao mesmo tempo, de modo que o verbo precisa ter um objeto, direto ou indireto. O fenômeno recebe o nome de pronome reflexivo na GT.

⁵Mais discussão sobre esse tipo de construção pode ser encontrada em [Cançado and Amaral 2010].

[Bechara 2012] indica que o pronome “se” como objeto exige um sujeito animado para o verbo, pois somente dessa forma o sujeito será agente e paciente da ação ao mesmo tempo. Em uma frase como “João se banha”, o sujeito é um ser animado e funciona como agente e paciente da ação do verbo. Já na frase “O banco só se abre às 10 horas”, para o autor, o sujeito inanimado impede a ocorrência de pronome reflexivo, sendo um caso de voz passiva. No caso de “Ele se chama João”, sabe-se que, embora animado, o sujeito não é agente da ação, restando a anotação de verbo pronominal.

No PetroGold não foram encontradas ocorrências de “se” como pronome reflexivo (objeto direto ou indireto). Para confirmar que a inexistência do pronome reflexivo no PetroGold está correta, verificamos todos os sujeitos de verbos a que se associam o pronome “se”. A análise dos 281 lemas não retornou nenhum sujeito animado, o que justificaria a ausência do pronome reflexivo no *corpus*, sugerindo ser uma característica dos textos do domínio a frequência baixa ou nula de frases em que o sujeito é animado, um dos requisitos elencados por [Bechara 2012].

4. Resultados

O corpus PetroGold v3 conta com 1.960 ocorrências de “se”, sendo 75 conjunções subordinativas e 1.885 pronomes expletivos, conforme tabela 1. Todos os 1.885 usos expletivos passaram por revisão, uma vez que precisaram ter sua etiqueta modificada para se acrescentar a informação relativa à indeterminação do sujeito, passivização ou uso pronominal do verbo.

<i>expl:impers</i>	<i>expl:pass</i>	<i>expl:pv</i>	total de pronomes
278	807	800	1.885

Tabela 1. Frequência das classificações do pronome “se”

Distribuímos os pronomes “se” expletivos pelos verbos aos quais se associam, e alguns dos verbos podem aparecer associados a mais de um tipo de pronome “se” expletivo (os exemplos 4 a 7 a seguir exemplificam esses casos). Como se vê na tabela 2, 25 verbos podem se associar a pronomes “se” de dois tipos diferentes. Desses, 22 (88%) ora são usados pronominalmente e ora na voz passiva sintética, enquanto somente 3 (12%) são usados ora como verbo pronominal, ora como oração de sujeito indeterminado. Nenhum dos verbos se associa a três categorias ao mesmo tempo, e não encontramos no corpus nenhum verbo que se associe ora a um pronome “se” que indique voz passiva sintética, ora índice de indeterminação do sujeito.

deprel do “se”	nº de verbos
<i>expl:impers</i>	21
<i>expl:pass</i>	154
<i>expl:pv</i>	142
duas categorias	25
três categorias	0

Tabela 2. Número de verbos que se associam aos tipos de pronome “se”

O maior número de verbos que podem ser utilizados tanto na voz passiva sintética como na forma pronominal já era esperado – [Azeredo 2000], por exemplo, comenta

sobre o fenômeno da cristalização do “se” em verbos de voz passiva, os quais, pela frequência de uso, vão se tornando pronominais. Sintaticamente, o fato de que muitos verbos podem ser empregados das duas formas é explicável pela semelhança estrutural – ambas requerem um sujeito sintático na frase, sendo que a diferenciação é realizada semanticamente ao interpretar se o sujeito é paciente do conteúdo verbal ou não. Por exemplo, nas frases 4 e 5, o verbo em destaque é “ajustar”. O primeiro, porém, tem como sujeito o substantivo “modelos”, sendo que, na interpretação dos autores, a frase não permite inferir que haveria um agente, propositalmente omitido da oração, responsável por ter ajustado os modelos no contexto em que o verbo foi utilizado, diferentemente da segunda oração, onde um agente não identificado ajustou a “frequência”, que é sujeito paciente da oração.

4. **expl:pv:** A partir do coeficiente de correlação, percebe-se que todos os modelos **se ajustaram**.

5. **expl:pass:** A bomba de água foi acionada com uma frequência de 30 Hz e então **ajustou-se** a frequência baseando-se na vazão de água desejada.

Já a diferenciação entre orações com sujeito indeterminado e uso pronominal do verbo pode ser explicada em termos puramente sintáticos. Nos exemplos 6 e 7, o verbo em destaque é “chamar”. Na primeira frase, o pronome relativo “que” retoma “pasta oleosa”, sujeito da oração relativa cujo núcleo é “se chama” e cujo complemento verbal é “petróleo”. Por tratar-se de uma oração com sujeito sintático, a anotação é a de verbo pronominal, diferentemente da segunda frase, em que “bentos” é objeto direto (sem a possibilidade da leitura como sujeito paciente) e “organismos” objeto indireto, faltando um sujeito sintático para a oração (que tampouco está elíptico), marcando, portanto, indeterminação do sujeito⁶.

6. **expl:pv:** Admite-se que o petróleo foi formado há milhões de anos pelo acúmulo de diferentes seres vivos como a decomposição de plânctons - seres que são geralmente encontrados na zona costeira, mares, oceanos e estuários - esses seres teriam se acumulados no fundo dos mares, rios e lagos e soterrados pela ação do movimento da crosta terrestre e posteriormente com o passar dos anos transformando-se em uma pasta oleosa que hoje **se chama** petróleo (VAZ, 2011).

7. **expl:impers:** Em biologia marinha e limnologia, **chama-se** bentos aos organismos que vivem no substrato, fixos ou não, em contraposição com os pelágicos, que vivem livremente na coluna de água.

A ausência de verbos compartilhando o –se na voz passiva sintética e o –se como indeterminação do sujeito pode ser explicada pelo fato de que os dois fenômenos são muito distintos sintaticamente, sendo que no primeiro há um sujeito sintático na oração, que é marcada por um VTD ou VTDI, e no segundo fenômeno não há sujeito, sendo utilizado um verbo VI ou VTI.

Os resultados da avaliação intrínseca do modelo gerado utilizando o *dataset* após as revisões do pronome –se como material de treino podem ser verificados na tabela 3. Entre parênteses, é possível verificar a variação, em pontos percentuais, das métricas de avaliação quando comparadas às métricas anteriores às revisões.

⁶A análise seria diferente caso “organismos” não fosse preposicionado: “Os organismos se chamam bentos” (verbo pronominal).

UPOS	LEMMA	LAS
98,40% (+0,05 p.p.)	98,46% (-0,09 p.p.)	88,71% (-0,39 p.p.)

Tabela 3. Avaliação intrínseca após as revisões do pronome “se”

A métrica relativa ao aprendizado de classe gramatical foi a única que obteve uma melhora (+0.05 p.p.). Isso pode ser explicado por alguns motivos: (a) foram realizadas correções sistemáticas relativas a quando o “se” é pronome ou conjunção subordinativa, facilitando o aprendizado automático de POS; (b) foi realizada uma simplificação das informações morfológicas do pronome expletivo “se” – originalmente, eram anotados como tendo atributos morfológicos de um pronome de terceira pessoa, como se fosse um objeto, portanto herdando as características do objeto da oração. Como se trata de um pronome expletivo, que não representa nem um sujeito nem um objeto, removemos completamente suas informações morfológicas, o que pode ter facilitado o aprendizado do etiquetador.

O decréscimo de 0,09 p.p. na avaliação de lematização, por sua vez, pode ser explicado pelo fato de que havia 22 palavras “sudeste” ou “Sergipe”, abreviadas como “SE”, mas que tinham o lema anotado como “se”, em letras minúsculas, o que facilitava o aprendizado uma vez que, independentemente de a palavra estar em caixa alta ou não, o lema era sempre o mesmo. Quando desfizemos essa anotação de lema, diferenciando o “se” pronome do “SE” abreviação para “sudeste” e “Sergipe”, introduzimos um pequeno obstáculo que pode ter refletido no decréscimo.

Já a métrica que diz respeito ao aprendizado de dependências (LAS) teve o desempenho piorado em 0,39 ponto percentual. Esse dado pode ser explicado por termos introduzido uma granularidade previamente inexistente no corpus quando adicionamos três novas classes para o pronome “se” – *expl:impers*, *expl:pass* e *expl:pv*. Antes das revisões, a classe *expl* obtinha 100% de acertos pois era a única para todos os casos de pronome “se”. Nessa nova versão, os resultados de acerto para as três novas classes foram de, respectivamente, 82,3%, 91,3% e 86,8%. Soma-se a isso o fato de que os verbos são polissêmicos e, dependendo do contexto, as orações – e portanto o “se” – podem ser interpretadas de uma forma ou de outra, conforme discutimos.

Além da anotação revista do *corpus* PetroGold v3, estamos disponibilizando também alguns recursos lexicais que podem ser úteis para outros projetos de anotação ou para alimentar tarefas linguístico-computacionais. Em um repositório dedicado a este trabalho⁷, disponibilizamos: (1) três listas com as frases que tiveram o pronome –*se* classificados como do tipo *expl:impers*, *expl:pass* ou *expl:pv*; (2) uma lista com todos os verbos associados ao pronome –*se*, organizados pela frequência com que o –*se* foi anotado usando as diferentes etiquetas e um exemplo de frase para cada tipo, e (3) a lista de verbos associados a pronomes –*se* (e as respectivas frases) em que os pronomes –*se* foram anotados de forma diferente, indicando polissemia do verbo, que admite mais de um fenômeno quando associado ao pronome. Os recursos funcionam como apêndices desse trabalho, e a ideia é que possam ser processados computacionalmente com facilidade.

⁷Disponível em: <https://github.com/alvelvis/recursos-lexicais-se>. Acesso em 22 de jun. 2023.

5. Considerações finais

Realizamos uma descrição linguística e discutimos os resultados do processo de anotação do pronome *–se* no *treebank* PetroGold (v3). O recurso integra o projeto *Universal Dependencies*, que prevê diferentes etiquetas para anotar o pronome *–se* mas não apresenta estruturas típicas da língua portuguesa, as quais anotamos tomando como base o estudo de gramáticas do português.

Como resultados, discriminamos as 1.960 ocorrências do “se” no *corpus* por classe sintática e apresentamos o número de verbos que se associam a cada um dos tipos do pronome *–se* expletivos. Com a revisão, foi possível perceber o comportamento de certos verbos com relação ao *–se*, o que por sua vez se reflete em uma constatação importante para o PLN: nem sempre um *corpus* mais bem anotado levará às melhores medidas de avaliação (medida F1). Em nosso caso, o fato de o *corpus* ter sido consistentemente revisto foi justamente o que fez o desempenho do modelo piorar.

Agradecimentos

Os autores agradecem ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, processo #130495/2021-2), à FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, processo #E-26/202.433/2022) e à ANP (Agência Nacional de Petróleo, Gás Natural e Biocombustíveis, Brasil, associada ao investimento de recursos oriundos das Cláusulas de P,D&I, por meio de Termo de Cooperação entre a Petrobras e a PUC-Rio) pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Referências

- Azeredo, J. C. d. (2000). Fundamentos de gramática do português. *Rio de Janeiro: Jorge Zahar*.
- Bagno, M. (2012). *Gramática pedagógica do português brasileiro*. Parábola Ed.
- Bechara, E. (2012). *Moderna gramática portuguesa*. Nova Fronteira.
- Cançado, M. and Amaral, L. (2010). Representação lexical de verbos incoativos e causativos no português brasileiro. *Revista da ABRALIN*, 9(2):123–147.
- Cunha, C. and Cintra, L. (2016). *Nova gramática do português contemporâneo*. LEXIKON Editora Digital Ltda.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.
- de Souza, E. (2023). *Construção e avaliação de um treebank padrão ouro*. Mestrado, PUC-Rio.
- 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.
- dos Santos Silva, A. (2021). O clítico se no português brasileiro como índice de indeterminação do sujeito. *EDUCTE: Revista Científica do Instituto Federal de Alagoas*, 12(1):1683 a 1692.

- Duran, M. S. and Aluísio, S. M. (2011). O tratamento da partícula “se” para fins de anotação de papéis semânticos. *II Jornada de Descrição do Português-Proceedings of 8th STIL-Cuiabá*, pages 24–26.
- Duran, M. S., Scarton, C., Aluísio, S., and Ramisch, C. (2013). Identifying Pronominal Verbs: Towards Automatic Disambiguation of the Clitic ‘se’ in Portuguese. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 93–100.
- Freitas, C. and de Souza, E. (2021). Sujeito oculto às claras: uma abordagem descritivo-computacional/Omitted subjects revealed: a quantitative-descriptive approach. *Revista de Estudos da Linguagem*, 29(2):1033–1058.
- Hartmann, N. S., Duran, M. S., and Aluisio, S. M. (2014). Filling the gap: inserting an artificial constituent where a subject is omitted in portuguese. In *WORKSHOP ON TOOLS AND RESOURCES FOR AUTOMATICALLY PROCESSING PORTUGUESE AND SPANISH (TORPOR), I, São Carlos, Proceedings [...]*. São Carlos: SBC.
- Lopes, E. M. and Namiuti-Temponi, C. (2017). A ordem e a função do clítico se no português clássico. *Entrepalavras*, 7(2):151–169.
- 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, 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.
- Vieira, C. B. and de Sá, T. M. M. (2015). Pronome apassivador? uma perspectiva cognitiva na análise do pronome se. *Palimpsesto-Revista do Programa de Pós-Graduação em Letras da UERJ*, 14(21):411–426.
- 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.

A funcionalidade dos adjetivos em dois gêneros discursivos: uma investigação com base nas dependências universais

André V. Lopes Coneglian¹, Adriana Pagano¹, Carlos Perini¹

¹Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
31.270-901 – Belo Horizonte – MG – Brazil
{coneglian, apagano, perini}@ufmg.br

Abstract. *This paper presents an investigation of adjectives in two discourse genres, Aesopian fables and essays on women's etiquette, drawing on annotations based on Universal Dependencies. The goal is to argue for a more diverse sample of texts with respect to discourse genre when building annotated treebanks. This is meant to allow for representing the different ways in which lexis and grammar are activated in different genres. As the results point out, there are important differences with respect to the use of adjectives in the two samples of texts, not only in terms of frequency of occurrence, but also, and more importantly, in terms of the type of dependency relations adjectives are implicated in.*

Resumo. *Este artigo apresenta uma investigação da ocorrência de adjetivos em dois gêneros discursivos, a fábula esopiana e colunas jornalísticas de comportamento feminino, baseada na anotação morfossintática de acordo com o do modelo das Dependências Universais. Objetiva-se apresentar evidências sobre a necessidade de se construir corpora anotados contemplando textos diversos, tendo-se vista que o acionamento do léxico e da gramática da língua varia nos diversos gêneros discursivos. Os resultados apontam uma diferença significativa no uso de palavras adjetivas em textos dos dois gêneros discursivos em questão, não só em termos de frequência de ocorrência, mas também no tipo de relação de dependência da qual participam.*

1. Introdução

Nos últimos anos, as Dependências Universais (*Universal Dependencies* – UD) [de Manerffe et al 2021] têm se estabelecido como um modelo de anotação eficiente para o Processamento de Língua Natural (PLN). Isso se deve, em parte, ao esforço coletivo de pesquisadores, em nível mundial, que têm construído bancos de dados anotados segundo esse modelo para línguas individuais. O estudo que ora se apresenta soma-se a esses esforços de construção de bancos de dados anotados. No entanto, o principal objetivo não é o desenvolvimento de bancos para tarefas em PLN, mas, sim, o uso de tais bancos na análise gramatical, neste caso, do português brasileiro. Está na base desta investigação o entendimento de que a construção de bancos de dados linguísticos anotados pode servir tanto ao processamento de língua natural quanto à descrição e análise linguística.

Mais especificamente, o que se busca mostrar neste trabalho é que o acionamento da gramática da língua é sensível ao gênero discursivo, isto é, que o sistema gramatical do português brasileiro é mobilizado de maneira diferente em diferentes gêneros discursivos. Esta hipótese é particularmente relevante no contexto do PLN, porque, historicamente, os grandes bancos de dados são constituídos, na sua maioria, de textos jornalísticos – apesar de que iniciativas recentes vêm integrando gêneros variados do português brasileiro [Souza et al, 2021; Pardo et al, 2021].

Nessa direção, este trabalho examina a ocorrência de adjetivos em dois gêneros discursivos, a fábula esopiana e a coluna de comportamento feminino. A ocorrência dos adjetivos é avaliada em termos dos tipos de relações de dependências (*deprels*) que tais palavras estabelecem. As *deprels* que o adjetivo estabelece são interpretadas segundo as funções sintáticas que o adjetivo pode desempenhar no português brasileiro [Neves, 2011, 2018] (Seção 2).

Assim, por meio do exame da ocorrência de adjetivos em textos de dois gêneros discursivos, objetiva-se apresentar evidências de que a diversidade de gêneros de uma língua contemplada nos corpora anotados enriquece a descrição linguística, uma vez que o acionamento do léxico e da gramática da língua varia nos diversos gêneros discursivos.

Surpreendentemente, há, na área do PLN, uma frente de investigação sobre a classe dos adjetivos bastante consolidada (por exemplo, Raskin e Nirenburg [1998]; Bouillon e Viegas [1999]; Kim e de Marneffe [2013]). Em geral, esses estudos concentram-se em tarefas de reconhecimento morfossintático e anotação de propriedades lógico-semânticas dos adjetivos. Este estudo, no entanto, como já se indicou, vai na direção de estabelecer indutivamente [Givón, 1995] quadros descritivos da gramática do português com o auxílio de recursos de PLN, neste caso, as UD's.

Este trabalho organiza-se em seis seções, sendo a primeira, esta Introdução. Na Seção 2, discute-se a categoria “adjetivo” na gramática do português, apresentando-se suas propriedades morfossintáticas e funcionais básicas, sob a perspectiva de uma gramática de base textual. Na Seção 3, discute-se o adjetivo no modelo das UD's, fazendo-se um equacionamento entre as relações de dependência (*deprels*) e as macrofunções dos adjetivos. Na Seção 4, apresentam-se a constituição da amostra textual que constitui o universo desta análise e os métodos para processamento automático e revisão dos textos segundo o modelo das UD's. Na Seção 5, apresentam-se os resultados, procedendo-se à sua discussão. A Seção 6 traz as considerações finais e possíveis extensões e aplicações desta pesquisa.

2. A categoria “adjetivo” na gramática do português

O debate linguístico a respeito da classe dos adjetivos tem sido notoriamente polarizado. Tanto há autores que afirmam que há línguas que não dispõem dessa classe em seu sistema gramatical, quanto há aqueles que, reconhecendo a existência dessa classe em uma língua, diluem-na nas classes de substantivo e verbo [cf. Dixon, 2004, 2010].

O português brasileiro é uma língua em que claramente se observa uma classe unificada de adjetivos. Do ponto de vista morfológico, há aquelas palavras que, de fato, são adjetivas, como *bonito* e *caro*, e também há palavras que se formam adjetivas por derivação morfológica, como *brilhante* e *cansado*. A distinção entre esses dois tipos de adjetivos é, em geral, descrita em obras gramaticais como uma diferença entre **adjetivos primitivos** e **adjetivos derivados** [Bechara, 2009; Neves, 2011, 2018]. Neste estudo, não consideramos a diferença entre essas duas classes morfológicas.

A segunda distinção que algumas gramáticas apontam, resulta nas subclasses de **adjetivos simples**, como *doentio* e *tristonho*, e de **locuções adjetivas**, como é o caso de *do inverno* e *de transporte*. As locuções adjetivas, no português, têm a forma de um sintagma preposicionado, formado de uma preposição mais um sintagma nominal. Para este estudo, são considerados apenas os adjetivos simples (Seção 4).

No que diz respeito às propriedades distribucionais dos adjetivos em português, Neves [2011, p. 180-184] documenta cinco funções sintáticas que podem ser desempenhadas por adjetivos, as quais são exemplificadas a seguir: (i) função **atributiva** (01), (ii) função **predicativa** (02), (iii) função **apositiva** (03), (iv) função **argumental** (04), (v) função **de substantivo** (05).

- (01) Uma criatura **alegre** predispõe sempre os outros à simpatia [...] (Coluna)
- (02) Sou **belo, alto**, de bom porte, e sou **útil** para tetos de templos e para navios. (Fábula)
- (03) [...] o escaravelho foi até ela, **faminto**, esmolar comida. (Fábula)
- (04) ... digo-o baseada na experiência, que adquiri sobre a arte de embelezar a mulher e atrair a atenção **masculina**. (Coluna) = atenção do homem
- (05) A fábula mostra que os que não se opõem à circunstância e aos **mais fortes** estão melhores do que os que competem com os **superiores**. (Fábula)

Como se discutirá mais adiante, são poucas as ocorrências em que o adjetivo tem a mesma distribuição de substantivos, na amostra deste estudo. Na Seção 4, a seguir, discutimos esses casos, mostrando-os como possivelmente problemáticos para a anotação das relações de dependência segundo o modelo das UD.

3. A categoria “ADJ” nas Dependências Universais

O modelo das UD apresenta um conjunto de etiquetas de classes de palavras universais (Universal Parts of Speech – UPOS). O foco deste trabalho está naquelas palavras que são anotadas como ADJ(etivo).

Segundo o modelo das UD, o adjetivo pode ser anotado em português como estabelecendo basicamente três relações: (i) *amod*, (ii) *xcomp* e (iii) *advcl*. Ele pode ser também a palavra que recebe a etiqueta *root*, quando é o núcleo de um predicado nominal. O Quadro 1, abaixo, traz as correspondências entre as funções sintáticas descritas na Seção 2 e as *deprels*.

Função sintática, segundo Neves (2011) <i>deprels</i> , segundo as UD	função atributiva	função argumental	função apositiva	função predicativa			função de substantivo
				root	advcl	xcomp	*

Quadro 1. Correspondências entre funções sintáticas dos adjetivos em português e suas respectivas *deprels*.

Muito notavelmente, a *deprel amod* compreende as funções atributiva, argumental e apositiva do adjetivo em português, ao passo que o adjetivo em função predicativa pode ser anotado como *xcomp* ou receber a etiqueta *root*. Adiante-se já que essas três etiquetas de *deprels* e as respectivas funções são as que têm com maior frequência de ocorrência na amostra deste estudo.

Tanto no caso das funções atributiva e argumental quanto na função predicativa, a correspondência com as *deprels* naturalmente decorre das propriedades distribucionais dos adjetivos em português. Como se viu nas ocorrências (01) e (04), acima, exemplos de adjetivos em função atributiva e argumental, respectivamente, nota-se que os adjetivos apresentam a mesma distribuição, seguindo o substantivo núcleo do SN de que fazem

parte. No que diz respeito à função predicativa, o adjetivo pode aparecer em duas distribuições absolutamente diferentes: como núcleo de um predicado nominal com verbo cópula, como ilustra (02), ou como predicado secundário (nos termos de van der Auwera e Malchukov [2005]), como ilustra (06).

(06) Para que esses olhares e essa admiração, porém, não se desviem decepcionados.
(Coluna)

A anotação de adjetivos que suscita ponderações diz respeito aos casos em que adjetivos assumem a função de substantivo, como ilustra (05), daí o * na célula correspondente na Figura 1. Podem-se apontar dois problemas a respeito dessa questão, um teórico e outro operacional. O problema de ordem teórica é, na verdade, a categorização de Neves (2011), que atribui ao adjetivo, nesta configuração em particular, função de substantivo. Na literatura tipológica, especialmente em Dryer (2007), casos como este são considerados como “sintagmas nominais sem núcleo”, mais particularmente, um sintagma nominal sem núcleo em que são realizados apenas determinantes e modificadores.

O problema operacional está na acomodação dessas perspectivas descritivas no modelo de anotação das UD. Ora, se for adotada a visão de Neves [2011], que documenta um verdadeiro deslizamento categorial de palavras adjetivas, pode ser difícil resolver tal desencontro categorial no modelo das UD, principalmente porque uma das diretrizes de anotação geral do modelo é que se dê preferência à anotação da palavra de acordo com sua classe morfológica – adjetivos são ADJ, substantivos são NOUN. A proposta alternativa de Dryer [2007] permite uma operacionalização mais facilmente alinhada às UD, talvez porque a proposta é tipológica por natureza. Porém, se adotarmos a proposta de Dryer [2007], será oportuno fazer o enriquecimento das anotações incorporando-se as dependências *enhanced* [cf. Nivre et al, 2018], porque o sintagma nominal sem núcleo é um caso de elisão de núcleo, caso este cuja anotação pode ser efetuada nas dependências *enhanced*.

De um modo geral, pode-se dizer que as correspondências entre funções sintáticas dos adjetivos em português e as *deprels* (Quadro 1) são correspondências prototípicas. Isso porque, como se verá nas análises na Seção 5, há adjetivos que estabelecem outras *deprels*, como *ccomp*, *acl:rel*, *advcl*, entre outras. No entanto, não há necessidade de contemplar essas correspondências no Quadro 1, porque, na verdade, se o adjetivo é anotado como estabelecendo alguma dessas *deprels*, o adjetivo é, na verdade, o núcleo de um predicado nominal. Veja-se como ilustração a ocorrência (07) e sua anotação na Figura 1. Na *deprel advcl* o adjetivo pode também ocorrer em estrutura de predicação secundária. Este caso se discute na Seção 5.

(07) Nunca diminuí-lo ou recriminá-lo porque não é brilhante, não é rico ou atraente.
(Coluna)

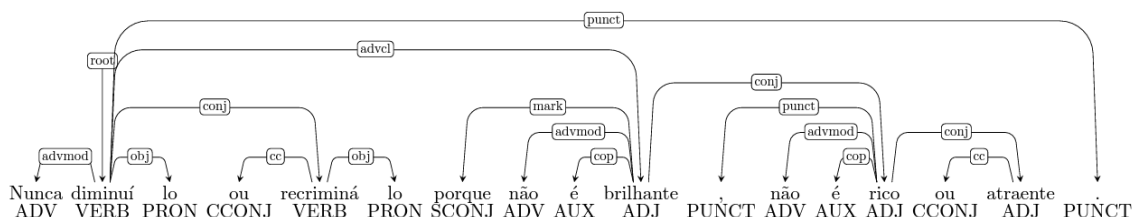


Figura 1. Anotação da ocorrência (07).

4. Material e métodos de análise

O *corpus* de análise é formado de textos dos gêneros discursivos fábula esopiana [Dezotti, 2018] e coluna jornalística de comportamento feminino [Nunes, 2008]. Para este trabalho, a escolha desses dois gêneros justifica-se pelo fato de que predominam neles diferentes sequências textuais, sendo a fábula esopiana um texto predominantemente narrativo e a coluna jornalística, predominantemente injuntivo e argumentativo. Essa diferença de predominância, por si, já garante a mobilização de diferentes recursos gramaticais na composição de textos desses gêneros [cf. Adam, 2019].

Os textos da amostra das fábulas esopianas foram traduzidos diretamente do grego [Dezotti, 2018]. Os textos da amostra de colunas jornalísticas de comportamento feminino são de autoria de Clarice Lispector e foram publicados originalmente nos extintos jornais *Correio da manhã* e *Diário da noite* [Nunes, 2008]. A Tabela 1 apresenta a composição das amostras.

Tabela 1. Composição da amostra de análise.

Gênero discursivo	Número de textos por amostra	Número de palavras na amostra
Fábula esopiana	64	7.909
Coluna jornalística	45	8.017

As colunas jornalísticas foram extraídas de um arquivo em formato pdf, convertidas para o formato txt codificação UTF8. As fábulas foram digitadas e convertidas para o formato txt codificação UTF8. Ambas as amostras foram revisadas para correção de potenciais problemas de conversão ou digitação. Para a anotação do *cópus*, foi utilizado o *framework* do projeto *Universal Dependencies v.2* (Nivre et al., 2020), que consiste em 17 etiquetas para anotação de classes gramaticais e 37 etiquetas de relações sintáticas, além de sub-relações. As amostras de texto foram primeiramente anotadas de forma automática por meio da ferramenta UDpipe¹ (Straka et al. 2016), com um modelo de língua portuguesa que utiliza o Bosque-UD v. 2.10 de textos jornalísticos (Rademaker et al. 2017), da qual resulta um arquivo em formato CONLL-U. Esse arquivo CONLL-U foi convertido em um arquivo formato csv para revisão da anotação automática,. A conversão do arquivo para formato csv foi feita por meio de um script Python, tipo *parser*, que recebe o arquivo CONLL-U como parâmetro de entrada e gera um arquivo no formato csv como saída. A anotação morfossintática e sintática segue as diretrizes do português brasileiro de Duran [2021, 2022].

5. Resultados e discussão

A análise que se apresenta nesta seção está pautada pelos seguintes parâmetros: a) comparação geral da ocorrência de adjetivos nas duas amostras de textos; b) verificação da função sintática do adjetivo em relação às *deprels* das UD.

No que diz respeito à ocorrência de palavras com a etiqueta ADJ em cada uma das amostras de texto, a Tabela 2 revela que essas palavras possuem uma frequência pouco superior ao dobro em colunas jornalísticas de comportamento feminino do que nas fábulas esopianas.

Tabela 2. Número de ocorrências de adjetivos na amostra de textos e sua frequência relativa.

Gênero discursivo	Número de ocorrência de adjetivos	Frequência relativa (N. adj/N. total de palavras)
Fábula esopiana	218	0,0276
Coluna jornalística	460	0,0573

¹ Disponível em: <https://lindat.mff.cuni.cz/services/udpipe/>.

Esse resultado representa um achado imprevisto. Esta investigação antecipava uma diferença no número de ocorrências dos adjetivos nos dois gêneros em questão, mas não se fazia previsão sobre em qual das duas amostras seria verificado o maior número de ocorrências de adjetivos. Para explicar essa diferença, no entanto, seria necessário ampliar o escopo da investigação, de modo a considerar fatores como finalidade discursiva, conteúdo temático e estilo – os quais, obviamente, não se enquadram no âmbito do estudo que se faz aqui.

A Tabela 3 traz a distribuição das palavras com a etiqueta ADJ quanto ao tipo de *deprel* que elas estabelecem nos enunciados em que ocorrem. É interessante notar que, tanto nas fábulas quanto nas colunas jornalísticas, predomina a *deprel amod*, que compreende, em português, os adjetivos em função atributiva e em função argumental. Em nenhuma das duas amostras, nenhuma outra *deprel* atinge uma frequência tão alta quanto *amod*.

Tabela 3. Distribuição de ADJ por etiqueta de *deprel* na amostra de textos.

<i>Deprel</i>	Gênero discursivo			
	Fábula esopiana		Coluna de comportamento	
	N. de ocorrências	Frequência relativa	N. de ocorrências	Frequência relativa
acl:relcl	5	0,0229	6	0,0130
advcl	7	0,0321	9	0,0195
amod	117	0,5366	276	0,6
ccomp	10	0,0458	15	0,0326
conj	21	0,0963	57	0,1239
csubj	0	0	4	0,0086
obj	3	0,0137	0	0
obl	2	0,0091	0	0
parataxis	0	0	7	0,0152
root	23	0,1055	29	0,0630
xcomp	30	0,1376	57	0,1239
Total	218	1	460	1

Pela distribuição das *deprels* nas amostras, pode-se perceber, também, que há: a) diferença significativa entre as ocorrências de adjetivo que recebem a etiqueta *root*, o que significa que, há maior ocorrência de estruturas de predicado nominal nas fábulas do que nas colunas jornalísticas; b) baixa ocorrência de adjetivos em *csubj*, (utilizado como recurso para construir uma definição nas colunas jornalísticas, como se vê em (08)); e c) ocorrência de *obj* e *obl*, exclusivamente nas fábulas, em estrutura de sintagmas nominais sem núcleo substantivo, como se vê em (09) e como se ilustrou em (05), na qual o adjetivo passa a ser o núcleo da relação de dependência.

(08) Ser **feminina** em doses maciças é pretender sedução por atacado. (Coluna)

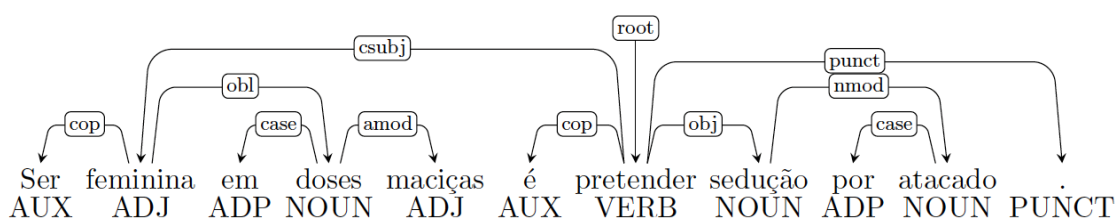


Figura 2. Anotação da ocorrência (08).²

² As imagens foram criadas com o pacote tikz-dependency em um editor de LaTeX.

(09) A fábula mostra que os empregados sentem saudades dos donos anteriores, sobretudo quando experimentam os novos. (Fábula)

No que diz respeito à ocorrência de adjetivos nas *deprels* *acl:relcl*, *advcl*, *ccomp*, *csubj*, *parataxis*, pode-se dizer que todos desempenham função predicativa na organização da estrutura dos enunciados. Observamos que não há, nem nas fábulas nem nas colunas, um predomínio da função predicativa dos adjetivos.

Merecem destaque, também, as ocorrências de *xcomp* que, no que diz respeito à classe adjetiva, podem corresponder a dois tipos distintos de estruturas: (i) uma em que o adjetivo funciona como predicado secundário (10), seja em relação ao sujeito, seja em relação ao objeto – essa distinção, no entanto, não é captada no modelo de anotação básico das UD's; e (ii) outra em que o adjetivo é o núcleo de uma construção resultativa, como (11).

(10) As moscas acham o azul francamente repelente. (Coluna)

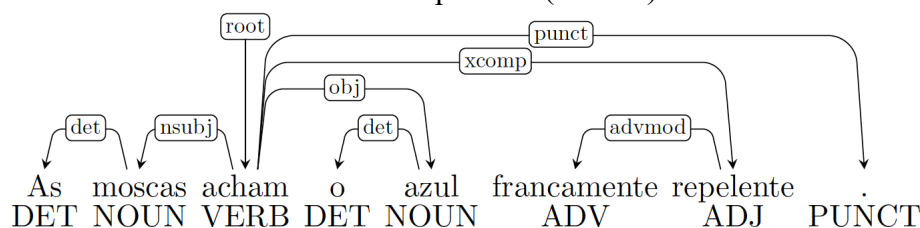


Figura 3. Anotação da ocorrência (10).

(11) Os açougues ficariam mais livres de moscas se pintassem portas e janelas de azul. (Coluna)

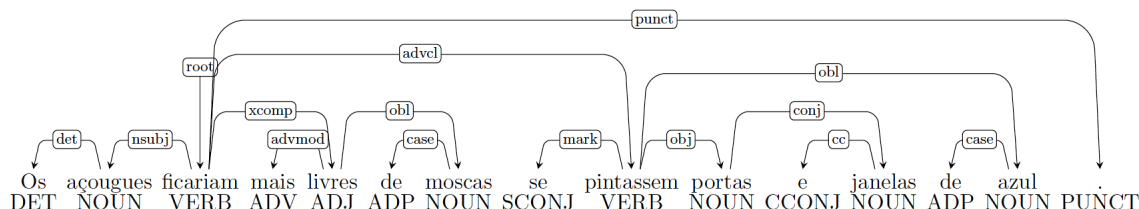


Figura 4. Anotação da ocorrência (11).

Comparando-se os diagramas de anotação de (10) e (11), nas Figuras 3 e 4, respectivamente, nota-se uma discrepância, também de ordem operacional. Em (10), o adjetivo *repelente* constitui um predicado secundário do *obj*, *azul*, mas a anotação da *deprel* de *xcomp* é com o verbo *achar*. Na anotação de (11), por outro lado, o adjetivo *livre* constrói uma predicação secundária resultativa do *nsubj*, *açougues*.

Finalmente, há um outro caso de predicação secundária no português, ilustrado em (12). Nesses casos, a diretriz de anotação das UD's recomenda que o adjetivo seja anotado como estabelecendo a *deprel* *advcl*. Veja-se a representação na Figura 6.

(12) Um leão que jazia doente em uma caverna disse à amigável raposa (Fábula)

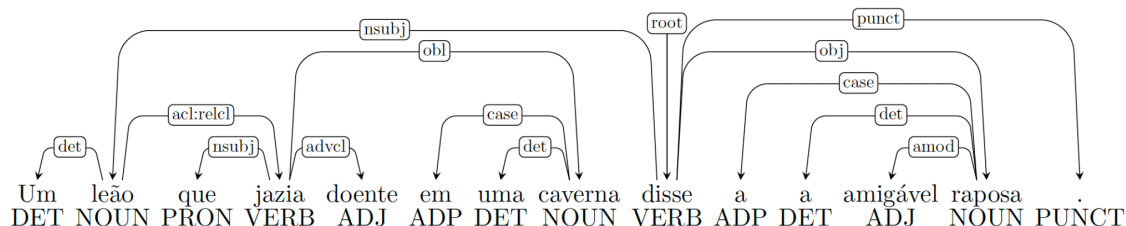


Figura 6. Anotação da ocorrência (12).

Ocorre que, em português, diferentemente do inglês, por exemplo, essa sentença tem uma interpretação ambígua: o adjetivo *doente* pode ter tanto orientação para o participante (o leão estava doente) quanto pode ter orientação para o evento (o leão jazia *doentemente). Tanto na amostra das fábulas quanto na das colunas, (12) foi a única ocorrência desse tipo de estrutura.

A análise apresentada aqui não pretende resolver esta questão, mas considerá-la para o estabelecimento de diretrizes de anotação. A pergunta que se faz, para a qual não se oferece resposta aqui é: Qual o melhor procedimento de anotação de estruturas de predicação secundária de modo a manter as relações entre termos dependentes, uma vez que nem sempre o predicado secundário liga-se ao predicado principal? Essa pergunta problematiza a diretriz geral de anotação das UD's, segundo a qual o *xcomp* deve ligar-se ao verbo. Em português – e em muitas outras línguas [van der Auwera e Malchukov, 2005; Croft, 2022] – o adjetivo, nessas estruturas, liga-se a um nome, como aqueles na posição de *nsubj* ou *obj*, nesses casos, e não ao verbo da predicação primária. Por aí se vê a dificuldade de conciliar, nas diretrizes de anotação, fatos particulares de cada língua com o modelo tipológico geral. Retorna-se a este ponto nas Considerações Finais.

6. Considerações finais

Este estudo trouxe apenas uma amostra de análise, para a classe dos adjetivos, com a problematização teórica em cuja base está a comparação da gramática em diferentes gêneros discursivos. O estudo revelou características distintivas do acionamento do adjetivo nos gêneros examinados e, a partir daí, discuti o potencial que o modelo das UD's para investigações descritivas do português. Mais abrangentemente, se se pretende comparar o acionamento da gramática em diferentes gêneros, faz-se necessário direcionar a análise para as propriedades globais, comparando-se a frequência de ocorrência de todas as UPOS e todas as *deprels*. Transcendendo a anotação morfossintática possibilitada pelas UD's, a expansão da análise dos dados deste estudo pode ser feita, para fins de enriquecimento desta descrição, por um lado, pela anotação das classes semânticas dos adjetivos [Croft, 2022; Dixon, 1982, 2004, 2010] e pela anotação da sua função textual-discursiva [Chafe, 1982; Thompson, 1988, 2000].

Um ponto de discussão que naturalmente decorre das análises é justamente a conciliação entre fatos descritivos de línguas particulares e a proposta geral do modelo. No que diz respeito ao adjetivo em português, não se pode propor, neste ponto das análises, uma sugestão de anotação que capte a distinção entre adjetivos em função atributiva e em função argumental, visto que a sua forma de expressão é a mesma na gramática do português (veja-se a discussão na Seção 2). Do ponto de vista teórico, o fenômeno precisa ser investigado mais sistematicamente para se determinar em que medida a função argumental do adjetivo decorre da composição semântica entre substantivo e adjetivo (como em *atenção masculina*). Do ponto de vista prático da anotação, essa distinção talvez possa ser resolvida na anotação das dependências *enhanced*. O mesmo se pode dizer para os casos em que os adjetivos estabelecem *deprel* de *xcomp* em estruturas de predicação secundária.

Espera-se ter sugerido que, no que diz respeito à construção de banco de dados anotados, são necessárias **diversidade de gêneros discursivos**, por um lado, de modo a garantir a diversidade nas estruturas gramaticais que compõem as amostras, bem como **a centralidade de uma teoria tipológica de gramática**, por outro, de modo que seja possível acomodar fatos particulares do português em contraponto com generalizações tipológicas.

Referências

- Adam, J-M. (2019) Textos: tipos e protótipos, Editora Contexto, São Paulo.
- Bechara, E. (2009) Moderna gramática portuguesa, 36ª edição, Nova Fronteira, Petrópolis.
- Bouillon, P. e E. Viegas. (1999) The description of adjectives for natural language processing: theoretical and applied perspectives. In *Atelier Thématique*, TALN, 1999.
- Chafe, W. (1982) “Integration and involvement in speaking, writing, and oral literature”, In: Spoken and written language: exploring orality and literacy, edited by D. Tannen, Ablex, New Jersey.
- Croft, W. (2022) Morphosyntax: constructions of the world’s languages, Cambridge University Press, Cambridge, UK.
- de Marneffe, M-C et al. (2021) Universal dependencies, In *Association for Computational Linguistics*, pages 255-308
- Dezotti, M. C. (2018) A tradição da fábula: de Esopo a La Fontaine, Editora Unesp, São Paulo.
- Dixon, R. M. W. (1982) Where have all adjectives gone?, Mouton de Gruyter, Berlin.
- Dixon, R. M. W. (2004) “Adjective classes in typological perspective”, In: Adjective classes: a cross-linguistic perspective, edited by R. M. W. Dixon e A. Aikhenvald, Oxford University Press, New York.
- Dixon, R. M. W. (2010) Basic linguistic theory, vol. 2, Oxford University Press, New York.
- Dryer, M. S. (2007) “Noun phrase structure”, In: Language typology and syntactic description, vol 2: complex constructions, edited by T. Shopen, Cambridge University Press, Cambridge, UK.
- Duran, M. S. (2021) Manual de anotação de PoS tags. Relatório Técnico, n. 434. NILC-ICMC/USP. Disponível em: <https://sites.google.com/icmc.usp.br/poetisa>. Acesso em 29 de junho de 2023.
- Duran, M. S. (2022) Manual de Anotação de Relações de Dependência: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em: <https://sites.google.com/icmc.usp.br/poetisa>. Acesso em 25 de junho de 2023.
- Givón, T. (1995) Functionalism and grammar, John Benjamins, Amsterdam.
- Kim, J-K, M-C de Marneffe. (2013). Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625-1630, Association for Computational Linguistics.
- Neves, M. H. M. (2011) Gramática de usos do português, 2ª edição, Editora Unesp, São Paulo.

- Neves, M. H. M. (2018) *A gramática do português revelada em textos*, Editora Unesp, São Paulo.
- Nivre, J et al. (2018) Enhancing Universal Dependency Treebanks: a case study. In *Proceedings of the Second Workshop on Universal Dependencies*, pages 102-107, ACL Anthology.
- Nivre, J. et al. (2020) Universal Dependencies v2: An ever growing multilingual treebank collection. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034-4043, European Language Resources Association.
- Nunes, A. (2008) *Só para mulheres – textos de Clarice Lispector*, Rocco, São Paulo.
- Pardo, T., M. Duran, L. Lopes, A. Felippo, N. Roman, M. Nunes, Maria. (2021). Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese. 1-10. 10.5753/stil.2021.17778.
- Raskin, V. e S. Nirenburg (1998). An applied ontological semantic microtheory of adjective meaning for natural language processing. In *Machine Translation*, vol. 13, pages 135-227, Springer.
- Rademaker, A. et al. (2017) Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 197–206, ACL Anthology.
- Souza, E., Silveira, A., Cavalcanti, T., Castro, M., & 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*, (pp. 29-38). Porto Alegre: SBC. doi:10.5753/stil.2021.17781
- Straka, M., J. Hajič, J. Straková. (2016) Udpipeline: trainable pipeline for processing conllu files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4290–4297, ACL Anthology.
- Thompson, S. (1988) “A discourse approach to the cross-linguistic category ‘adjective’”, In: *Exploring language universals*, edited by J. Hawkins, Blackwell, New York.
- Thompson, S. (2000) “Property concepts”, In: *Morphology: an international handbook of inflection and word-formation*, edited by G. Booji et al, Mouton de Gruyter, Berlin.
- van der Auwera, J. e A. Malchukov (2005) “A semantic map for depictive adjectivals”, In: *Secondary predication and adverbial modification: a typology of depictives*, edited by N. P. Himmelmann e E. Schultze-Berndt, Oxford University Press, New York.

Indução Gramatical para o Português: a Contribuição da Informação Mútua para Descoberta de Relações de Dependência

Diego Pedro Gonçalves da Silva¹, Thiago Alexandre Salgueiro Pardo¹

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

diegopedro@usp.br, taspardo@icmc.usp.br

Resumo. *Indução gramatical é uma tarefa que busca aprender automaticamente estruturas sintáticas a partir de texto. Poucos trabalhos de indução gramatical foram produzidos direcionados para a língua portuguesa. Neste artigo, reproduzimos o trabalho de [Futrell et al. 2019] para a língua portuguesa e o estendemos ao incluir análise de informação mútua para relações sintáticas específicas. Utilizamos dois treebanks anotados e realizamos experimentos utilizando embeddings de dimensões variadas, demonstrando a hipótese de alta informação mútua para palavras em relações de dependência.*

1. Introdução

Na Linguística, sintaxe é definida como o estudo da organização das palavras (em termos de ordenação e estruturação) na formação de sentenças. Esse entendimento é compartilhado por diferentes visões sobre como a sintaxe deve ser formalizada [Chomsky 2014] [Bresnan et al. 2015]. Quase toda aplicação de Processamento de Línguas Naturais (PLN) necessita de algum conhecimento sintático para obter bons resultados, direta ou indiretamente codificados. Revisores gramaticais, sistemas de simplificação de textos e sistemas de extração de informação são algumas das aplicações que se beneficiam da representação explícita da sintaxe. As aplicações baseadas em grandes modelos de língua, por sua vez, acabam adquirindo noções de sintaxe em seu treinamento, mesmo que ela não seja completamente explicitada..

Dada a relevância da sintaxe, a Indução Gramatical (IG), também chamada de *parsing* não supervisionado [Klein e Manning 2004], é uma tarefa de interesse na comunidade de PLN. Apesar de ela ter a finalidade de induzir (“aprender”) automaticamente a gramática a partir de dados textuais sem anotações sintáticas [Klein e Manning 2004], vários autores realizam IG como tarefa semi-supervisionada (IGSS) ou supervisionada (IGS) [Headden III et al. 2009] [Spitkovsky et al. 2013]. Sendo assim, utilizaremos o termo (IGNS) para nos referirmos à tarefa de IG não supervisionada. É interessante notar que, independentemente de aplicações computacionais de PLN, a IGNS pode auxiliar em várias frentes. Na Linguística, pode ser útil para aprender a gramática de línguas mortas ou com escassez de recursos (como as indígenas) [Dahl et al. 2023]. Em Psicolinguística, pode ser utilizada para propor modelos de aquisição da linguagem [Bannard et al. 2009]. Em Bioinformática, IGNS é utilizada para inferir estruturas de DNA desconhecidas ou difíceis de serem encontradas em grandes bases de dados [Unold et al. 2020].

A sintaxe (e, por consequência, a IGNS) se vale de duas visões diferentes de representação: a gramática de constituinte e a gramática de dependência. A primeira

estuda como as sentenças são formadas por blocos básicos (sintagmas). No modelo de representação de gramática de dependência, foco desse artigo, estabelecem-se relações de dependências diretamente entre as palavras. A Figura 1 apresenta relações de sujeito (nsubj) (entre o verbo “chutou” e a palavra “menino”) e objeto (obj) (entre o verbo e a palavra “bola”), por exemplo.

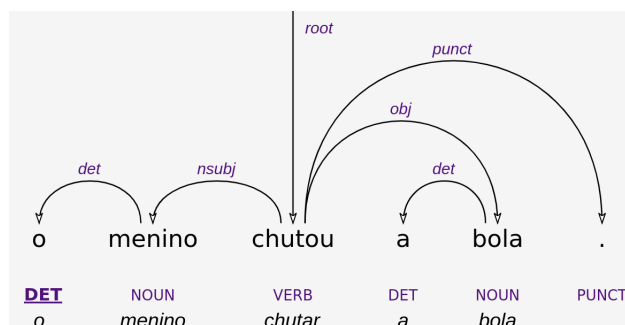


Figura 1. Exemplo de análise de dependência

Devido aos avanços da IGNS nos anos 2000, presumia-se que a IGNS se aproximaria da IGS em desempenho em breve, mas isso ainda não ocorreu [Bod 2007] [Lin et al. 2022]. A maioria dos trabalhos desenvolvidos nas últimas décadas utiliza algum tipo de anotação. Induzir gramática sem nenhuma informação prévia é uma tarefa bastante difícil. Outro desafio é a indução de gramática em sentenças independentemente do tamanho. A maioria dos trabalhos publicados utiliza sentenças de até 10 palavras. Por fim, as diferenças sintáticas entre as línguas apresentam mais um desafio que dificulta a padronização de técnicas para diferentes línguas. Por exemplo, as línguas chinesa, persa e tupi apresentam características linguísticas diferentes por fazerem parte de famílias de línguas diferentes [Theodor e Siebert-Cole 2020], o que pode dificultar a criação de um modelo unificado. No melhor de nosso conhecimento, não encontramos trabalhos publicados de IGNS específico para a língua portuguesa. O trabalho mais similar encontrado foi produzido por [da Costa e Kepler 2014], que implementa uma abordagem semi-supervisionada baseado no trabalho de [Klein e Manning 2004].

Nesse contexto, nosso objetivo neste artigo é explorar a tarefa de IGNS para o português. Em específico, focamos na reprodução de um experimento recente de uso da conhecida medida de Informação Mútua (IM) para tentar predizer palavras que possam estar relacionadas sintaticamente. A IM é uma medida de dependência, assim, quanto maior a informação mútua entre duas palavras, maior a chance de elas estarem relacionadas. Baseamo-nos no trabalho de [Futrell et al. 2019], que, usando IM aplicada a um corpus de milhões de palavras em inglês, mostrou que há uma maior IM entre palavras que mantêm relação de dependência do que entre palavras que não mantêm. Além de avaliar tal técnica para o português, vamos além e verificamos seu comportamento para relações específicas. Realizamos nossos experimentos com dados de *treebanks* alinhados ao modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021], amplamente adotado.

Na Seção 2, apresentamos brevemente os principais trabalhos relacionados. Em seguida, na Seção 3, descrevemos a abordagem aplicada no nosso estudo. Na Seção 4, apresentamos os resultados do estudo. Fazemos algumas considerações finais na Seção 5.

2. Trabalhos relacionados

Ao longo das últimas décadas, várias abordagens foram utilizadas em IGNS. A maioria dos trabalhos utilizam a abordagem gerativa, principalmente no uso do algoritmo *Expectation–maximization – EM* [Baker 1979], que é utilizado para estimar a probabilidade de variáveis não observáveis (árvores sintáticas em IGNS). Nos últimos anos, a modelagem neural vem ganhando bastante espaço.

Ao longo das duas últimas décadas, o modelo DMV (*Dependency Model with Valence*) [Klein e Manning 2004] exerceu grande influência para gramática de dependência. A ideia por trás do modelo DMV está no controle de geração da árvore sintática, que, para cada ramo a ser gerado (relação de dependência), utiliza-se de distribuições de probabilidade para tomar decisões de quando gerar ($P_{STOP}(\neg STOP|h, dir, adj)$) e qual ramo gerar ($P_{CHOOSE}(a|h, dir)$). As variáveis h , dir , a e dij são respectivamente a cabeça da relação, a direção em que o argumento será gerado (direita ou esquerda), o argumento a ser gerado e se o argumento já foi gerado na árvore na direção dir . O DMV é um dos vários modelos que utilizam o EM. Este foi o primeiro trabalho a ultrapassar o *baseline* de ramificação direita (*right-branching*) [Headden III et al. 2009], sendo bastante utilizado, mesmo com quase duas décadas de existência [Yang et al. 2020].

Muitos trabalhos foram influenciados por [Klein e Manning 2004]. Um dos mais relevantes, [Headden III et al. 2009] estendeu o modelo DMV para aplicar uma abordagem Bayesiana, em vez de EM, utilizando uma gramática lexicalizada (cada nó da árvore sintática contém também informação sobre o léxico a que se refere). [Cohen e Smith 2009] optou por substituir a distribuição *Dirichlet* pela Logística, pois, apesar de a primeira ser mais fácil de treinar, ela não permite um meio explícito de forma flexível para calcular a covariância entre dois eventos, conforme descreve [Blei e Lafferty 2005]. O trabalho alcançou 42% de *Direct Dependency Accuracy – DDA* (quando considera a direção de geração da árvore sintática) no corpus WSJ ∞ , para sentenças de qualquer tamanho.

O trabalho de [Spitkovsky et al. 2010] obteve bons resultados a partir da aplicação de *curriculum learning* [Bengio et al. 2009], que inicia o treinamento com dados menos complexos e aumenta a complexidade dos dados até que toda a base de dados tenha sido utilizada. A mudança de complexidade contribui para que se reduzam as chances de cair em máximos locais (um dos problemas no uso de EM usado para problemas não convexos). Este trabalho obteve 45% de DDA no WSJ ∞ . Mais recentemente, [Han et al. 2019b] propôs o *Lexicalized Neural Dependency Model with Valence* (L-NDMV), um modelo lexicalizado que utiliza DMV com redes neurais. Esse trabalho constatou que, ao explorar características lexicais, a tarefa de IGNS ganha em desempenho. O L-NDMV foi o primeiro trabalho a ultrapassar a marca dos 60% de DDA no WSJ ∞ , enquanto que os trabalhos supervisionados ultrapassam a marca dos 95% em *Unlabeled Attachment Score – UAS* (quando não considera a direção de geração da árvore) [Lin et al. 2022].

[Yang et al. 2020] atingiu o estado da arte ao construir o modelo probabilístico com mais de um nível de distância de hierarquia (além de *filhos*, *pais* e *irmãos* também considera *avôs*, *netos* e *tios*, por exemplo) entre os nós da árvore. Outros trabalhos atingiram o estado da arte ao estender o modelo DMV com redes neurais [Han et al. 2019a] [Han et al. 2017] [Jiang et al. 2016]. Todos estes trabalhos utilizam algum tipo de informação léxica e redes neurais para contribuir com o desempenho. [Shen et al. 2021] usa o conceito de distância e altura sintática para segmentar a sentença em partes menores.

[Drozdov et al. 2019] aplica o algoritmo *Inside-Outside* – *IO*, que pode ser visto como uma instância do EM, em redes neurais.

Todos os trabalhos citados utilizam algum tipo de anotação no treinamento. Recentemente, [Pate e Johnson 2016] treinou o modelo DMV com milhões de palavras para induzir dependência sem uso de anotação. Uma vez que o modelo não utiliza categorias morfossintáticas, é utilizada inferência Bayesiana aplicada a gramáticas livres de contexto probabilísticas. Apesar de já existirem estudos anteriores que utilizavam apenas palavras como entrada para o modelo [Seginer 2007], estes eram apenas para constituintes.

O uso de IM, em específico, é algo que vem sendo relativamente pouco explorado, apesar de seu claro apelo para a tarefa. [Magerman e Marcus 1990] foi o primeiro trabalho a aplicar IM em IGS, mas foi recentemente que IM começou a ser aplicado em tarefas de IGNS. Em um trabalho recente, [Futrell et al. 2019] constatou que pares de palavras que têm relação sintática apresentam uma maior IM quando comparados a pares de palavras sem relação. Esta hipótese também foi aplicada em indução gramatical por [Hoover et al. 2021], que utilizou o modelo de língua pré-treinado para calcular informação mútua entre palavras considerando o contexto.

A seguir, detalhamos o método de [Futrell et al. 2019] e como o reproduzimos.

3. Método de indução gramatical

O trabalho de [Futrell et al. 2019] utilizou um corpus com 320 milhões de *tokens* anotados automaticamente. Os autores analisam três variáveis: palavras (*words*), categorias morfossintáticas (*pos*) e grupos lexicais (*lex*). A última variável é resultante de agrupamento. O trabalho propôs um agrupamento com os 60K *tokens* mais frequentes, incluindo *stopwords* e pontuação, a fim de ter uma dimensionalidade menor. Utilizando os vetores de 300 dimensões do modelo *Glove* para cada uma das 60K palavras, o trabalho agrupa os tokens em 300 grupos. Por exemplo, os *tokens* “carro”, “carros”, “automóvel” e “automóveis” fazem parte do mesmo grupo lexical. Para representar a variável *lex*, cada *token* no corpus é substituído pelo número do seu respectivo grupo.

A IM é calculada entre pares de palavras, categorias morfossintáticas e grupos lexicais. Para pares de palavras, por exemplo, na sentença “O menino chutou a bola”, alguns dos pares possíveis são <o,chutou> e <chutou,bola>. O primeiro par não tem relação de dependência (indicada no experimento como *nondep*). O segundo par tem relação, conforme apresentado na Figura 1 (indicada como *dep*). [Futrell et al. 2019] quis também saber o desempenho de pares aleatórios. Ele descreveu estes pares como *permuted*. O mesmo é estabelecido para categorias morfossintáticas e grupos lexicais. Ao todo, para cada variável, 3 experimentos são realizados. O cálculo da IM ocorre entre o termo cabeça da relação *h* e o dependente *d*, cuja fórmula é apresentada abaixo. A fórmula calcula a probabilidade de haver uma relação entre duas variáveis, que podem ser palavras, categorias morfossintáticas, grupos lexicais e relações sintáticas (usadas neste trabalho).

$$IM = \log \frac{P(h, d)}{P(h)P(d)}$$

Para avaliar seu modelo, [Futrell et al. 2019] utiliza dois *baselines*: pares permutados (*words perm*, *lex perm*, *pos perm*) e pares sem relação de dependência (*words nondep*,

lex nondep, *pos nondep*). [Futrell et al. 2019] utilizou estes dois baselines porque o primeiro considera uma relação aleatória na sentença, podendo ser de dependência (como em <o,menino>) ou não (<o,a>). Assim, o *baseline* permutado deve apresentar melhor desempenho do que o *baseline* sem dependência se existir maior informação mútua entre relações de dependência do que não dependência. Estes *baselines* são comparados com os resultados das 3 variáveis para relação de dependência (*words dep*, *lex dep*, *pos dep*).

Diferentemente de [Futrell et al. 2019], utilizamos dois corpora anotados por humanos, disponíveis na página do projeto UD: Bosque [Afonso et al. 2002] e Petrogold [de Souza et al. 2021]. Para o Bosque, as sentenças anotadas com Português do Brasil correspondem a 90K *tokens* e 4.205 sentenças originárias do CETENFolha [Linguateca 2023], construído com textos jornalísticos. O corpus PetroGold é formado por 19 teses e dissertações na área de óleo e gás, constituído por 232k *tokens* e 9k sentenças. O corpus Bosque contém 4,16% das sentenças com até 3 *tokens* e 8,3% das sentenças com mais de 40 *tokens*. O PetroGold contém 4,5% das sentenças com até 3 palavras e 24,7% com sentenças acima de 40 *tokens*. Além da distribuição diferente de tamanho de sentença, constatamos que os corpora apresentam também diferenças na distribuição de categorias morfosintáticas e funções sintáticas, apesar de essas diferenças serem pequenas.

Adotamos o método de [Futrell et al. 2019], mas utilizamos apenas anotações que foram produzidas por humanos, sem análise automática. Além disso, não selecionamos os *tokens* mais frequentes, uma vez que o tamanho do vocabulário dos corpora compilados é menor que o tamanho proposto por [Futrell et al. 2019]. Em vez disso, apenas utilizamos todo o vocabulário dos corpora que é representado no modelo Glove treinado para a língua portuguesa [Hartmann et al. 2017], totalizando um vocabulário de 21.428 palavras.

Neste trabalho, utilizamos *embeddings* de 50, 300 e 600 dimensões com base na análise realizada por [Hartmann et al. 2017] para gerar os grupos lexicais. Para definir os 300 grupos, [Futrell et al. 2019] utiliza uma matriz de similaridade. No entanto, [Futrell et al. 2019] não informa como essa matriz de similaridade foi construída. Deduzimos que a matriz foi construída usando similaridade de cosseno entre as *embeddings* de cada *token*. Devido às limitações computacionais, foram utilizadas apenas duas casas decimais para representar os valores na matriz de similaridade.

Além dos dois corpora isoladamente, também usamos a combinação deles. Para cada corpus, foram realizadas 3 execuções, uma para cada uma das 3 dimensões, contabilizando um total de 9 execuções. O código utilizado para a realização dos experimentos foi o mesmo disponibilizado por [Futrell et al. 2019].

4. Resultados

Na Figura 2, são apresentados os gráficos com os resultados para IM utilizando todos os corpora. Resolvemos utilizar os nomes originais das variáveis utilizados no trabalho do [Futrell et al. 2019] para facilitar a reprodução do estudo. Incluímos a variável *fs*, que representa as relações de dependência. Para o agrupamento, variável *lex*, foram utilizadas *embeddings* de 300 dimensões.

Na Figura 2(a), observa-se que, conforme o número de pares aumenta, decresce a informação mútua em todos os grupos, sem tendência de convergência para um valor específico. Este mesmo comportamento é observado no trabalho de [Futrell et al. 2019]

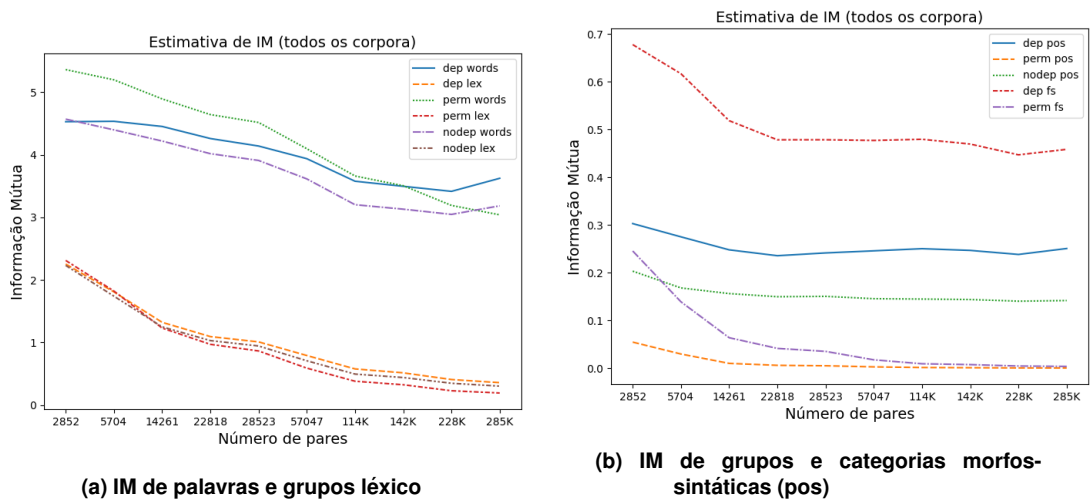


Figura 2. IM por número de pares usando 300 dimensões com todos os corpora

para a língua inglesa, mesmo utilizando um corpus dezenas de vezes maior que o nosso. Isso sugere que a IM entre pares que contenham relações de dependência pode seguir o mesmo padrão para diferentes línguas. Na Figura 2(b), para as variáveis *pos*, observa-se uma estabilidade, também apresentando comportamento similar ao trabalho de [Futrell et al. 2019]. Observa-se também uma IM maior para relações sintáticas do que para categorias morfossintáticas, mesmo com maior esparsidade no grupo de relações sintáticas.

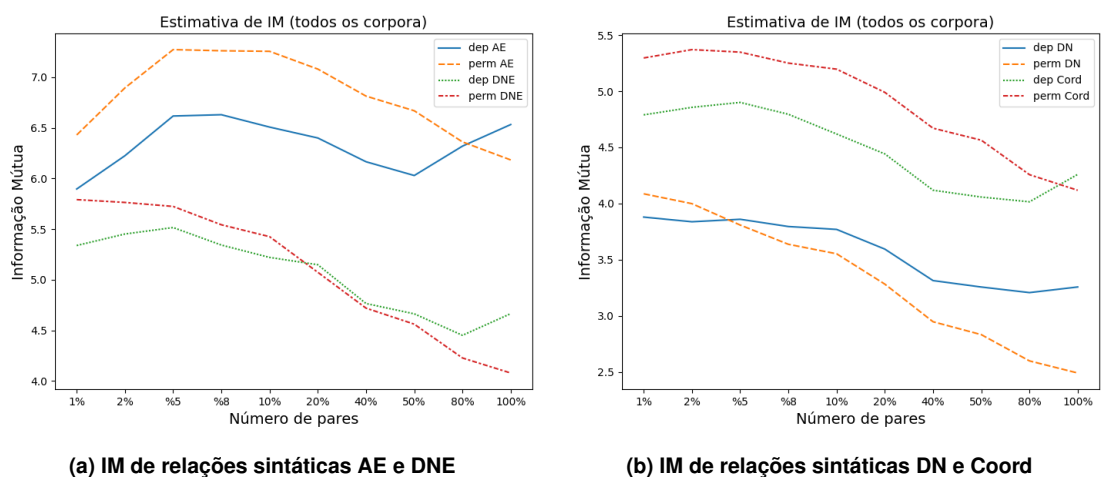


Figura 3. IM por porcentagem dos pares utilizados com todos os corpora

Na Figura 3, são apresentados os resultados para IM de relações sintáticas separadas pelos grupos definidos pela UD: Argumentos Essenciais (AE) (*Core arguments*, que incluem, por exemplo, as relações sintáticas mais importantes da sentença, como sujeito e objetos), Dependentes Não Essenciais (DNE) (*Non-core dependents*, que incluem, por exemplo, relações de vocativo, modificadores adverbiais e verbos auxiliares), Dependentes Nominais (DN) (*Nominal dependents*, que incluem, por exemplo, modificadores de substantivos e de adjetivos) e coordenações (Coord) (*Coordination*, que incluem, por exemplo, fenômenos variados, como a coordenação por conjunções, expressões multipa-

lavra e relações especiais). Na Figura 3, podemos observar que, dos quatro grupos, AE apresenta a maior IM, assim como também é o único que não apresenta uma tendência de queda conforme o número de pares aumenta.

Durante os experimentos, observamos que o número de dimensões influencia na IM. Nos experimentos usando agrupamento, percebemos que, quanto maior o número de dimensões, menor será a IM, apesar de constatarmos um pequeno aumento de pouco mais de 5% da IM utilizando *embeddings* de 600 dimensões em comparação com *embeddings* de 300 dimensões. Não temos certeza do que pode ter causado esta variação, mas, uma vez que ocorreu uma redução de 50% de IM entre os grupos que utilizaram *embeddings* de 50 e 300, acreditamos que o resultado pode ser devido à alguma característica intrínseca aos *embeddings* utilizados. Não conseguimos identificar uma relação entre o tamanho da sentença e a IM. Um resumo dos resultados para cada corpus é apresentado na Tabela 1. Os dados sugerem que, quanto menor o número de pares, maior a informação mútua.

Tabela 1. Resumo das execuções para todos os corpora

Corpora	$\mu(\sigma)$ pa- lavras / sentença	Número de pares	IM dep words	IM nondep words	IM dep lex	IM non- dep lex
Bosque	21,5 (13,51)	70.938	4,638	4,016	0,352	0,301
PetroGold	30,0 (19,5)	221.987	3,275	2,893	0,213	0,187

Finalmente, realizamos experimentos usando as relações sintáticas (Tabela 2). Devido à diferença no número de pares entre as relações, realizamos os experimentos considerando 4 das relações mais relevantes: *nsubj*, *obj*, *iobj* e *xcomp*.

Tabela 2. Experimento utilizando informações sintáticas

Relações	Número de pares	IM dep words	IM permuted	IM dep fs	σ fs
<i>nsubj</i>	55.412	7,488	0,014	0,478	0,0057
<i>obj</i>	34.692	7,295	0,020	0,631	0,0061
<i>iobj</i>	677	5,026	0,520	1,232	0,0159
<i>xcomp</i>	8.357	5,897	0,088	0,940	0,0122

Os resultados apresentados na Tabela 2 demonstram que há uma diferença muito grande de IM entre relações de dependência e IM com relações permutadas. A relação *iobj* apresenta o melhor desempenho entre as demais relações quando se observa a direção da relação (IM de 1,232), provavelmente devido ao número pequeno de exemplos analisados, uma vez que a IM permutada é bastante alta e o desvio padrão também. No caso da relação *nsubj*, a terceira coluna representa a IM das palavras que fazem parte dessa relação de dependência (sem considerar a direção da relação). Percebe-se uma alta IM nesta categoria, provavelmente devido às características sintáticas do *nsubj*. Como ilustração de pares com alta IM, as maiores IM encontradas para *obj* foram para os pares **computadores – comprei** (com valor 0,03197), **anos – há** (0,01870) e **-se – trata** (0,01336).

5. Considerações finais

Reproduzimos o trabalho de [Futrell et al. 2019] usando corpora da língua portuguesa. Apesar de o tamanho dos corpora usados por [Futrell et al. 2019] e os usados neste trabalho serem bem diferentes, constatamos tendência de comportamento similares, com algumas pequenas diferenças, sugerindo que existe um padrão de comportamento mesmo em línguas pertencentes à famílias linguísticas diferentes. Diferentemente do estudo publicado por [Futrell et al. 2019], que anotou milhões de palavras automaticamente, utilizamos apenas anotações sintáticas de referência produzidas por humanos. Essas variações tornam inconclusivas comparações diretas entre os trabalhos. Além disso, [Futrell et al. 2019] não informa como foi construída a matriz de similaridades.

Trabalhos futuros incluem aplicar este experimento a outros corpora anotados, como o Porttinarí [Pardo et al. 2021], e realizar um estudo mais aprofundado sobre a influência nos resultados dos corpora, assim como das *embeddings* utilizadas.

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

- Afonso, S., Bick, E., Haber, R., e Santos, D. (2002). Floresta sinta(c)tica: A treebank for portuguese. In the *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 1698–1703.
- Baker, J. K. (1979). Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 132–132.
- Bannard, C., Lieven, E., e Tomasello, M. (2009). Modeling children’s early grammatical knowledge. In the *Proceedings of the National Academy of Sciences (PNAS)*, 17284–17289.
- Bengio, Y., Louradour, J., Collobert, R., e Weston, J. (2009). Curriculum learning. In the *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 41–48.
- Blei, D. M. e Lafferty, J. D. (2005). Correlated topic models. In the *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 147–154.
- Bod, R. (2007). Is the end of supervised parsing in sight? In the *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 400–407.
- Bresnan, J., Asudeh, A., Toivonen, I., e Wechsler, S. (2015). Lexical-functional syntax. John Wiley & Sons.
- Chomsky, N. (2014). Aspects of the Theory of Syntax, volume 11. MIT press.
- Cohen, S. B. e Smith, N. A. (2009). Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, 74–82.

- da Costa, P. B. e Kepler, F. N. (2014). Semi-supervised parsing of portuguese. In the *Proceedings of the Computational Processing of the Portuguese Language - 11th International Conference (PROPOR)*, 102–107.
- Dahl, V., Bel-Enguix, G., Tirado, V., e Miralles, J. E. (2023). Grammar induction for under-resourced languages: The case of ch’ol. In the *Proceedings of the Analysis, Verification and Transformation for Declarative Programming and Intelligent Systems - Essays Dedicated to Manuel Hermenegildo on the Occasion of His 60th Birthday*, 113–132.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., e Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 255–308.
- de Souza, E., Silveira, A., Cavalcanti, T., Castro, M. C., e Freitas, C. (2021). Petrogold corpus padrão ouro para o domínio do petroleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 29–38.
- Drozdo, A., Verga, P., Yadav, M., Iyyer, M., e McCallum, A. (2019). Unsupervised latent tree induction with deep inside-outside recursive autoencoders. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, 1129–1141.
- Klein, D. e Manning, C. D. (2002). A generative constituent-context model for improved grammar induction. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 128–135.
- Futrell, R., Qian, P., Gibson, E., Fedorenko, E., e Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. In the *Proceedings of the fifth international conference on dependency linguistics (depling)*, 3–13.
- Han, W., Jiang, Y., e Tu, K. (2017). Dependency grammar induction with neural lexicalization and big training data. In the *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1683–1688.
- Han, W., Jiang, Y., e Tu, K. (2019a). Enhancing unsupervised generative dependency parser with contextual information. In the *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, 5315–5325.
- Han, W., Jiang, Y., e Tu, K. (2019b). Lexicalized neural unsupervised dependency parsing. *Neurocomputing*, 105–115.
- Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J. S., e Alu’isio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In the *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 122–131.
- Hoover, J. L., Du, W., Sordoni, A., e O’Donnell, T. J. (2021). Linguistic dependencies and statistical dependence. In the *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2941–2963.
- Headden III, W. P., Johnson, M., e McClosky, D. (2009). Improving unsupervised dependency parsing with richer contexts and smoothing. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, 101–109.
- Jiang, Y., Han, W., e Tu, K. (2016). Unsupervised neural dependency parsing. In the *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 763–771.
- Klein, D. e Manning, C. D. (2004). Corpus-based induction of syntactic structure:

- Models of dependency and constituency. In the *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 478–485.
- Lin, B., Yao, Z., Shi, J., Cao, S., Tang, B., Li, S., Luo, Y., Li, J., e Hou, L. (2022). Dependency parsing via sequence generation. *Findings of the Association for Computational Linguistics*, 7339–7353.
- Linguateca (2023). Cetem publico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Linguateca, <http://www.linguateca.pt/CETENFolha/>, última visita: Junho de 2023.
- Magerman, D. M. e Marcus, M. a P. (1990). Parsing a natural language using mutual information statistics. In the *Proceedings of the 8th National Conference on Artificial Intelligence (AAAI)*, 984–989.
- Pardo, T. A. S., Duran, M. S., Lopes, L., Felippo, A. d., Roman, N. T., e Nunes, M. d. G. V. (2021). Portinari: a large multi-genre treebank for brazilian portuguese. In the *Proceedings of the XIII Symposium in Information and Human Language (STIL)*, 1–10.
- Pate, J. K. e Johnson, M. (2016). Grammar induction from (lots of) words alone. In the *Proceedings of 26th International Conference on Computational Linguistics (COLING)*, 23–32.
- Seginer, Y. (2007). Fast unsupervised incremental parsing. In the *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 384–391.
- Shen, Y., Tay, Y., Zheng, C., Bahri, D., Metzler, D., e Courville, A. C. (2021). Structformer: Joint unsupervised induction of dependency and constituency structure from masked language modeling. In the *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJNLP)*, 7196–7209.
- Spitkovsky, V. I., Alshawi, H., e Jurafsky, D. (2010). From baby steps to leapfrog: How “less is more” in unsupervised dependency parsing. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, 751–759.
- Spitkovsky, V. I., Alshawi, H., e Jurafsky, D. (2013). Breaking out of local optima with count transforms and model recombination: A study in grammar induction. In the *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1983–1995.
- Stevenson, A. e Cordy, J. R. (2014). A survey of grammatical inference in software engineering. *Science of Computer Programming*, 444–459.
- Theodor, C. C. e Siebert-Cole, E. (2020). Family tree of languages. <https://www.researchgate.net/publication/342850691> TREES of LANGUAGES 2022, última visita:junho 2023.
- Unold, O., Gabor, M., e Dyrka, W. (2020). Unsupervised grammar induction for revealing the internal structure of protein sequence motifs. In the *Proceedings of Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine (AIME)*, 299–309.
- Yang, S., Jiang, Y., Han, W., e Tu, K. (2020). Second-order unsupervised neural dependency parsing. In the *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 3911–3924.

Características gerais das leis federais brasileiras: um estudo exploratório sob o enfoque da Análise Multidimensional

Carolina Godoi de Faria Marques, Lúcia de Almeida Ferrari

Faculdade de Letras - Universidade Federal de Minas Gerais

carol.godoi@outlook.com.br, ferrari.lu@gmail.com

Abstract. *Legal language is a rich and multifaceted specialized language used in various contexts, such as in courts, laws, and law textbooks [Goźdz-Roszkowski 2012; Carapinha 2018]. Despite its richness, there is a paucity of studies describing the linguistic features of this language in Brazilian Portuguese. Aiming to contribute to this field, we investigated its use in Brazilian federal statutory legal texts analyzing their linguistic variation and verifying if they are a register [Biber and Conrad 2009]. To do so, we conducted, supported by Corpus Linguistics, a Multidimensional Additive Analysis [Biber 1988] of the LEX-BR-Ius corpus [Ferrari and Marques in compilation] and confirmed our hypothesis.*

Resumo. *A linguagem jurídica é uma linguagem especializada rica e multifacetada utilizada nos mais diversos contextos, tais como em tribunais, leis e livros didáticos de direito [Goźdz-Roszkowski 2012; Carapinha 2018]. Apesar da sua riqueza, há poucos estudos que descrevem as características linguísticas dessa linguagem no português brasileiro. No intuito de contribuir com tal campo escolhemos investigar os textos legais federais brasileiros analisando sua variação linguística e verificando se eles são um registro [Biber e Conrad 2009]. Para tanto, realizamos, guiados pela Linguística de Corpus, uma Análise Multidimensional Aditiva [Biber 1988] do corpus LEX-BR-Ius [Ferrari e Marques em preparação e 2022] e confirmamos nossa hipótese.*

1. Introdução

Este trabalho apresenta os resultados de uma pesquisa de mestrado que realizou uma Análise Multidimensional aditiva [Biber 1988] utilizando os dados do *LEX-BR-Ius* [Ferrari e Marques em preparação e 2022] e se insere no projeto homônimo que busca descrever linguisticamente os textos legais federais brasileiros. Iremos aqui detalhar as particularidades de tais textos sob o enfoque da Análise Multidimensional [Biber 1988], uma abordagem empírico-metodológica que se vale de métodos quantitativos para identificar e medir as características linguísticas de um dado corpus, calculando a variação nos textos que o compõem e interpretando seus resultados de forma qualitativa. Graças a tal metodologia, foi possível obter um panorama da linguagem utilizada nesses textos e verificar nossa hipótese de que os textos legais são um registro segundo a perspectiva de Biber e Conrad (2009).

A linguagem jurídica se insere no rol das linguagens especializadas por apresentar léxico e gramática específicos. É importante lembrar que sob tal especificação há uma variedade muito grande de situações, e, portanto, linguagens diversas, que vão desde textos escritos, como as normas jurídicas e as várias etapas processuais, a interações orais,

como os depoimentos e oitivas entre muitos outros [Tiersma 1999; Coulthard e Johnson 2007, 2010; Goźdz-Roszkowski 2012; Carapinha 2018; Lorz 2019]. Delimitar e descrever a linguagem jurídica é, portanto, tarefa árdua e longa que requer recortes específicos. Acreditamos estar contribuindo para tal através dessa pesquisa que se configura como uma etapa preliminar na busca de uma descrição detalhada da língua empregada nas normas legais brasileiras.

O trabalho de Berber Sardinha, Kauffmann e Acunzo (2014) descreveu, valendo-se da Análise Multidimensional do Corpus Brasileiro de Variação e Registro (CBVR) [Berber Sardinha, Kauffmann e Acunzo 2014], as características linguísticas de uma grande variedade de registros do português brasileiro (doravante PB), traçando desta forma um retrato da variação linguística no PB. O trabalho que apresentamos somou o *LEX-BR-Ius* às dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014), comparando-o com o CBVR e buscando evidenciar as diferenças entre eles e identificar as características próprias do nosso corpus.

Como dito acima, a linguagem jurídica representa um universo multifacetado que merece ser analisado em cada uma de suas especificidades. Assim como uma poesia é muito diferente de um romance ou de um conto ou de um artigo de jornal, normas legais são tipos textuais muito distintos de julgamentos, petições, sentenças e assim por diante. Mostraremos, nesta contribuição, que não somente as normas legais são um registro específico, mas também que os critérios de seleção para a sua análise resultam em aspectos linguísticos específicos.

2. Referenciais teóricos e metodológicos

Um corpus é um conjunto de textos autênticos, coletados de acordo com critérios específicos de arquitetura para que seja representativo de uma língua ou de uma determinada variedade linguística, armazenado e submetido a tratamento computacional, possibilitando buscas automáticas ou semiautomáticas [Sinclair 2004; McEnery e Wilson 2001; Tognini-Bonelli 2001; Baker, Hardie e McEnery 2006].

Dentre os métodos de análise de corpus temos a Análise Multidimensional (AMD) [Biber 1988] utilizada para o estudo da variação linguística. A AMD permite tanto o estudo sincrônico quanto diacrônico de quaisquer línguas. Para tanto, esse método adota a perspectiva do registro, definido por Biber e Conrad (2009) como uma variedade da língua com traços situacionais, linguísticos e funcionais específicos, utilizada em contextos comunicativos próprios. Conforme Biber (1988), a AMD tem como pressuposto que a co-ocorrência de traços linguísticos em determinados padrões é fruto de fatores funcionais a eles inerentes. A identificação desses padrões em textos de determinado registro e comparação com outros registros possibilita sua caracterização. Para a descrição da sua variação, os registros são distribuídos segundo seus traços linguísticos em dimensões. Um registro tem várias dimensões intrínsecas, reveladas a partir da submissão dos dados de um corpus a uma série de operações estatísticas e seus resultados a uma interpretação funcional. [Biber 1988; Biber e Conrad 2009; Berber Sardinha, 2010]

O termo variação se refere aqui às diferenças entre a linguagem natural utilizada na comunicação, escrita e falada, de informantes diversos na mesma situação e contexto de uso ou pelo mesmo informante em diafasias distintas [Biber e Conrad 2009]. A AMD evidencia quais traços morfológicos e sintáticos do registro predominam em cada

dimensão selecionada pelo pesquisador possibilitando cruzar tais informações e traçando, de tal maneira, um quadro preciso das características linguísticas do registro em análise [Biber 1988].

Segundo Biber (1988) e Berber Sardinha (2013a), ao adotar a AMD como método para estudar a variação linguística devemos escolher entre realizar a Análise Multidimensional completa (*Full MD*) ou a Análise Multidimensional aditiva (Aplicação de dimensões). Embora ambas forneçam dados linguísticos significativos que permitem o estudo da variação linguística nos registros analisados, elas se diferenciam pela profundidade e finalidade da análise, assim como pela metodologia empregada, se tratando assim de abordagens complementares.

A primeira identifica as características linguísticas dos registros em análise e estabelece suas dimensões de variação, fornecendo uma descrição precisa e detalhada seja dos registros que das dimensões. Para tanto, várias etapas metodológicas quantitativas e qualitativas devem ser seguidas. Esse tipo de AMD, entretanto, é limitada em escopo, uma vez que abrange uma quantidade finita de registros e seu rigor metodológico, especialmente em relação às exigências na compilação do corpus a ser analisado, e a complexidade dos cálculos estatísticos requeridos muitas vezes representa um impeditivo para a sua realização. [Biber 1988; Biber e Conrad 2009; Berber Sardinha 2013a]

Já a segunda incorpora os registros analisados a dimensões já identificadas por uma AMD completa, enriquecendo-a, pois aumenta seu escopo. Esses registros são então comparados com aqueles utilizados para identificar as dimensões às quais o corpus foi adicionado e o resultado é um panorama das características linguísticas e da variação dos registros estudados ao longo das dimensões. Para isso, o corpus utilizado deve ser compilado seguindo os mesmos parâmetros do corpus utilizado no estudo que identificou as dimensões adotadas. É importante ressaltar também que cabe ao pesquisador escolher em quais das dimensões adicionará seu corpus, podendo adicioná-lo a todas ou apenas algumas. A realização da AMD aditiva é considerada mais simples, flexível e rápida, pois compreende menos etapas e procedimentos estatísticos, sendo eles menos complexos, quando comparada à primeira, e tem um rigor metodológico menor. [Biber 1988; Berber Sardinha 2013a]

3. Materiais e métodos

Para realizar a pesquisa utilizamos dados extraídos do corpus *LEX-BR-Ius*, um corpus sincrônico de normas legais federais brasileiras que apresenta, entre suas escolhas metodológicas de representatividade: (a) textos inteiros de maneira a garantir sua organização interna e conteúdo [Sinclair 2004; Biber 1993; Ferrari e Marques 2022; Marques 2023]; b) textos selecionados com base em sua frequência de uso no mundo legal [Barbera e Onesti 2009; Ferrari e Marques 2022; Marques 2023]; normas legais em vigor no momento de sua extração [Ferrari e Marques 2022; Marques 2023]. O corpus é dividido em seções segundo o tipo de norma, seguindo a classificação adotada pelo Planalto Brasileiro: Constituição, Códigos, Estatutos, Emendas à Constituição, Leis complementares e Leis ordinárias.

Com o objetivo de identificar os traços linguísticos característicos dos textos legais federais brasileiros e determinar se eles podem ser considerados um registro submetemos nosso corpus a uma Análise Multidimensional Aditiva com o auxílio dos softwares: Microsoft Excel e “IBM SPSS Statistics 23” [IBM 2015]. Esclarecemos que,

para atender às exigências da AMD aditiva o corpus foi etiquetado morfossintaticamente e lematizado com o o PALAVRAS [Bick 2000 e 2014], o mesmo etiquetador utilizado no estudo que identificou as dimensões às quais adicionamos o corpus. Foi necessário também fazer um recorte nos dados, sendo utilizadas apenas as seções: Constituição, Códigos e Estatutos, de forma a evitar possíveis vieses e permitir uma comparação a mais fidedigna possível dos nossos dados àqueles do estudo anterior.

A análise foi precedida de um levantamento preliminar das pesquisas sobre linguagem jurídica, com foco naquelas que compreendiam os textos legais. Para a realização da AMD aditiva em si, optamos por utilizar os dados da Análise Multidimensional do CBVR adicionando nosso corpus a todas as dimensões de variação do português brasileiro por ela identificadas, quais sejam: (1) *Oral versus literate discourse*, (2) *Argumentation*, (3) *Involved versus informational production*, (4) *Directive discourse*, (5) *Future versus past time orientation* e (6) *Reported discourse*. Para tanto, primeiramente individualizamos os traços linguísticos presentes em cada uma dessas dimensões a partir dos dados fornecidos em Berber Sardinha, Kauffmann e Acunzo (2014). Em seguida, submetemos nosso corpus anotado a um tratamento no pós-processador “PALAVRAS Tag count” [Berber Sardinha 2013b] que realiza a contagem automática das ocorrências dos traços em cada texto do corpus.

A fim de equilibrar os dados e evitar vieses, normalizamos por mil palavras as ocorrências dos traços contabilizadas anteriormente e calculamos seus Z-escores. Posteriormente, calculamos o escore de dimensão de cada texto que compõe o corpus ao longo de cada dimensão analisada a fim de obter a carga fatorial de cada um deles. Já para obter as cargas fatoriais do corpus em cada dimensão, de forma a observar em que ponto da escala de cada uma delas ele se localiza calculamos a média dos escores de dimensão. Por fim, adicionamos os dados obtidos à planilha das médias de dimensão do CBVR a nós fornecida pelos autores do estudo e comparamos nossos dados àqueles. Para complementar nossa análise, submetemos os dados aos testes estatísticos ANOVA e R^2 e calculamos o *p-value* para verificar o grau de variação abarcado pelo corpus em cada dimensão e se ele se diferencia dos registros do CBVR de forma estatisticamente relevante.

4. Resultados

Conforme apresentado na seção anterior, comparamos nossos dados com o CBVR em todas as dimensões do PB. Nosso corpus obteve pontuações únicas em todas elas, entretanto, as pontuações obtidas nas dimensões 2 (*Argumentation*), 4 (*Directive discourse*) e 6 (*Reported discourse*) giraram em torno de 0. Diante desses resultados concluímos que a argumentação (dimensão 2) não exerce um papel fundamental nos textos legais, não sendo eles marcados pela presença substantiva de traços argumentativos nem pela sua ausência. Da mesma forma, os textos não são marcados pela instrucionalidade nem pela sua falta (dimensão 4). Identificamos, no nosso corpus, a presença de alguns traços linguísticos geralmente associados a ordens, instruções ou orientações para a execução de determinadas tarefas, mas eles apresentam pontuações tanto positivas quanto negativas baixas. Por fim, também o discurso indireto (dimensão 6) não é característico no nosso corpus. Embora alguns de seus traços estejam presentes nos textos legais sua pontuação não é suficiente para caracterizá-los como marcados pelo discurso indireto nem pela sua ausência. Provavelmente, como os textos legais são

impositivos, criados com o objetivo de informar e descrever as normas que regem os sujeitos de direito da forma mais clara e detalhada possível para evitar que hajam interpretações diversas da pretendida, a argumentação e a instrucionalidade exercem um papel secundário no seu texto, sendo o discurso direto preterido àquele indireto.

Por esta razão, aprofundaremos a seguir somente as dimensões mais relevantes para a caracterização dos nossos dados, ou seja, aquelas nas quais houve uma maior pontuação, quais sejam: 1, 3 e 5 e reproduziremos apenas parte dos dados referentes aos registros do CBVR.

4.1. Dimensão 1: *Oral versus literate discourse*

Essa dimensão abarca as características do discurso oral e do discurso letrado: ao longo do polo positivo estão os registros típicos do discurso oral e daquele negativo aqueles típicos do discurso letrado.

A seguir reproduzimos o gráfico dessa dimensão com a adição de nosso corpus.



Figura 1. Dimensão 1

Tanto nosso corpus quanto a seção “*Legislation*” do CBVR se localizam no polo negativo, sendo que a seção “*Legislation*”, pontuou de forma mais negativa que o nosso corpus, o que significa que, apesar de ambos serem discursos letrados, os textos dessa seção do CBVR têm uma maior ocorrência de traços típicos desse discurso que aqueles

do *LEX-BR-Ius*. Os registros desse polo são marcados pela alta densidade informacional, discurso formal e especializado. Dentre os traços típicos desse polo destacamos a forte presença de sintagmas nominais, substantivos compostos e abstratos, adjetivos, artigos definidos, preposições, participípios passados e nominalizações na posição de sujeito. A seguir trazemos exemplos dessas características nos nossos textos:

(i) Art. 1º Esta Lei institui o Estatuto da Igualdade Racial, destinado a garantir à população negra a efetivação da igualdade de oportunidades, a defesa dos direitos étnicos individuais, coletivos e difusos e o combate à discriminação e às demais formas de intolerância étnica. [Brasil 2010]

Nos exemplos acima reproduzidos temos uma presença expressiva de vários dos traços típicos dessa dimensão. Destacamos a presença de diversos sintagmas nominais de vários tamanhos, como: Esta Lei; o Estatuto da Igualdade Racial. Ressaltamos também o uso de vários substantivos abstratos como: discriminação e adjetivos como: negra, étnicos, etc. Citamos ainda o uso de artigos definidos e preposições simples e compostas. Temos também a presença de participípios passados como: destinado. Todos esses traços contribuem para a especificação, restrição e detalhamento do conteúdo veiculado explicam a alta densidade informacional e complexidade das estruturas sintáticas encontradas no nosso corpus.

4.2 Dimensão 3: *Involved versus informational production*

Nessa dimensão os registros estão dispostos segundo o nível de envolvimento/interação dos participantes no registro no polo positivo e seu caráter informativo no polo negativo, conforme pode ser observado no gráfico abaixo.

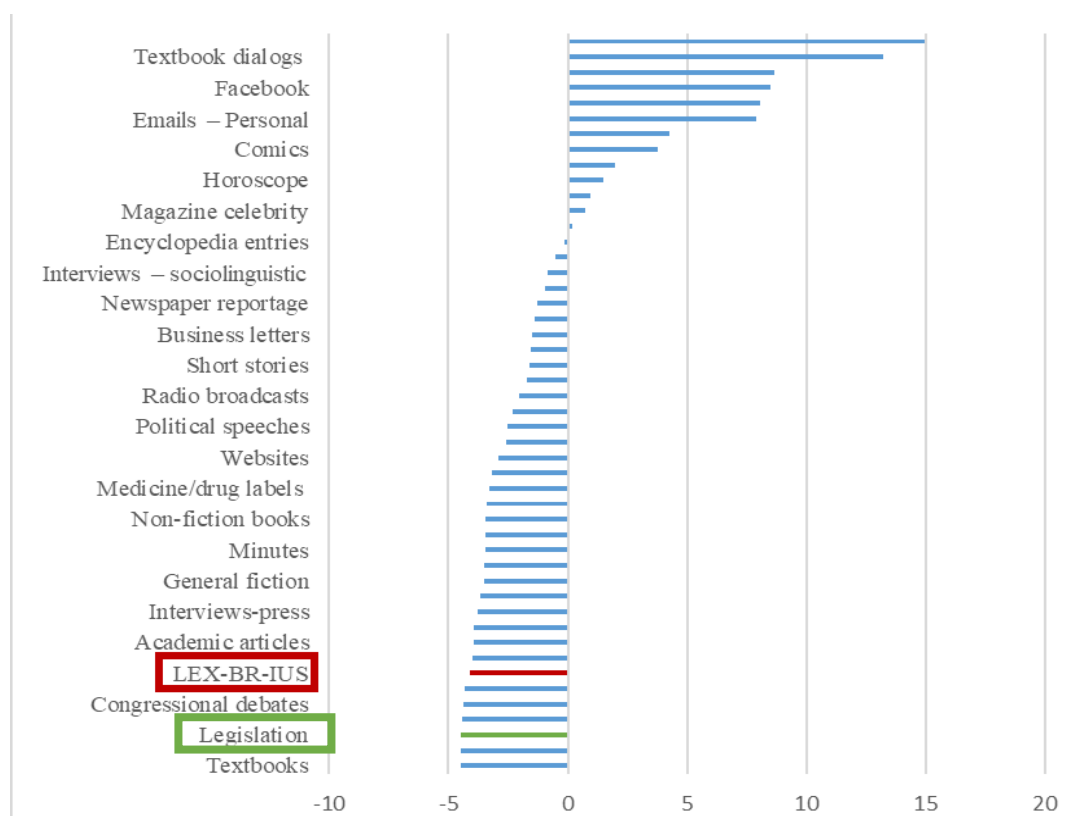


Figura 2. Dimensão 3

Como esperado, tanto nosso corpus quanto a seção “*Legislation*” pontuaram negativamente. Este polo apresenta apenas uma variável: relação *type-token*, que reflete a alta densidade lexical. Logo, seja nosso corpus que a seção “*Legislation*” apresentam pouca interação e alta carga informacional, tendo a última uma maior relação *type/token* que o primeiro. Como exemplo dessa característica nos nossos textos temos:

Art. 23. As ilhas ou ilhotas, que se formarem no álveo de uma corrente, pertencem ao domínio público, no caso das águas públicas, e ao domínio particular, no caso das águas comuns ou particulares. [Brasil 1934]

Neste exemplo podemos observar uma ampla gama de vocabulário especializado como: ilhotas, álveo, domínio público e domínio particular.

4.3 Dimensão 5: *Future versus past time orientation*

Na dimensão 5, por sua vez, temos, no polo positivo os registros cujo discurso é voltado para o futuro e, no polo negativo, aqueles voltados para o passado.

A seguir reproduzimos o gráfico dessa dimensão.

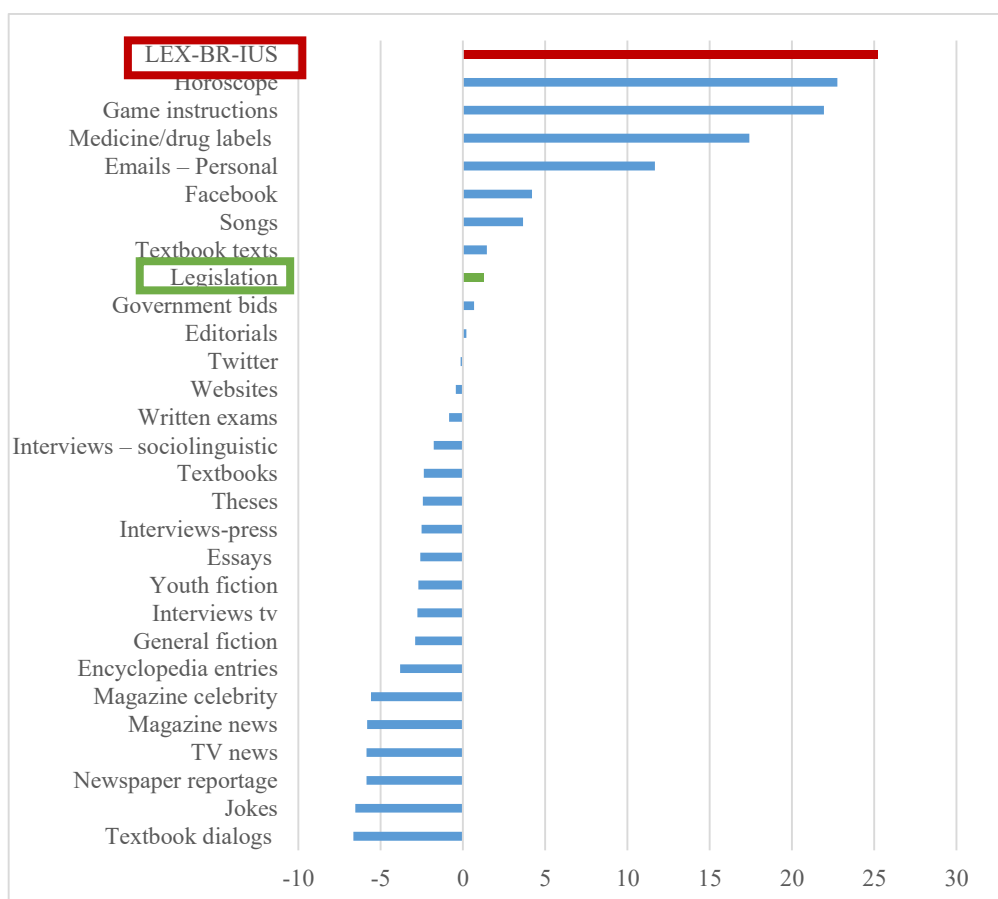


Figura 3. Dimensão 5

Essa é a dimensão na qual nosso corpus mais se destacou, pontuando superiormente a todos os registros do CBVR no polo positivo. Observa-se ainda, uma grande diferença entre a sua pontuação e aquela da seção “*Legislation*”, que é drasticamente inferior à nossa. O polo positivo é marcado por: alta frequência de verbos no futuro e uso dos modais *dever* e *poder*, conjunções coordenadas e orações subordinadas. São exemplos desses traços no nosso corpus:

Art. 8º § 2º O fornecedor deverá higienizar os equipamentos e utensílios utilizados no fornecimento de produtos ou serviços, ou colocados à disposição do consumidor, e informar, de maneira ostensiva e adequada, quando for o caso, sobre o risco de contaminação. [Brasil 1990]

No exemplo acima reproduzido observamos o uso do verbo modal *dever* no futuro e da conjunção coordenada “ou”, utilizados para prever e descrever os comportamentos exigidos dos destinatários da lei, assim como especificar e caracterizar os objetos desses comportamentos. De forma mais geral, o uso do futuro auxilia o legislador na descrição das normas, sua aplicação e consequências.

4.4 Significância estatística da variação

O resultado dos testes estatísticos empreendidos, quais sejam: ANOVA, p-value e R^2 indicam que nosso corpus se distingue significativamente dos registros do CVBR,, inclusive “*Legislation*” em todas as dimensões analisadas. Já em relação à variação abarcada pelos textos que compõem o nosso corpus, também essa é estatisticamente relevante em todas as dimensões, capturando entre 53.4% a 81.6% da variação prevista para as dimensões do PB.

5 Discussão

O estudo de Berber Sardinha, Kauffmann e Acunzo (2014) é o mais completo sobre a variação do português brasileiro já realizado até o momento. O corpus utilizado –CVBR– foi compilado sob medida para a sua realização visando abarcar o maior número possível de registros em PB de forma a obter uma descrição fidedigna da variação linguística nessa língua. Para tanto, o corpus amostra 48 registros orais e escritos, com 20 textos integrais por registro, totalizando 960 textos e aproximadamente 5.6 milhões de palavras. Dentre os registros que o compõe destacamos o “*Legislation*” composto por 20 textos legais que juntos totalizam 125 mil palavras (aprox. 2.2% do número de palavras do corpus). Tal seção é constituída por textos dos tipos normativos: Constituição, Códigos, Leis Ordinárias, Decretos, Medidas Provisórias e Atos Institucionais, sendo que alguns deles não são previstos pela atual Constituição (Atos Institucionais) e outros não estão mais em vigência. Não há também nenhuma separação interna na seção visando agrupar as diferentes espécies normativas. Cabe ressaltar que a maioria deles data de 2001, o que indicaria um recorte temporal, entretanto esta informação não consta da descrição do corpus. Da mesma forma, não foram disponibilizadas pelos autores informações importantes como os critérios adotados para a seleção dos textos, de onde foram extraídos ou quando, limitando-se os textos a terem dimensões similares àquelas das outras seções do CVBR.

A diferença, em alguns casos significativa, nos resultados obtidos entre a seção *Legislation* do CVBR e o *LEX-BR-Ius* parece-nos ser imputável ao nível de curadoria na seleção e limpeza dos textos que a compõem. O *LEX-BR-Ius* foi compilado respeitando o critério de atualidade e frequência de uso das normas legais [Ferrari e Marques 2022; Marques 2023], como apontamos acima, e com uma grande atenção na limpeza dos dados, conferida manualmente. Acreditamos que o registro “*Legislation*” não seja completamente representativo de todas as espécies normativas previstas na Constituição e inclui algumas não previstas, abarcando textos que não são mais aplicáveis e não tendo um critério de seleção claro. Além disso, da realização da limpeza não foram excluídas

as informações extratextuais como: índices, informações editoriais e números de página, o que pode enviesar os dados.

6 Conclusão

A AMD aditiva empreendida e seus resultados, juntamente com aqueles dos testes estatísticos, indicam que nosso corpus tem cargas fatoriais únicas e variação estatisticamente significativa em todas as dimensões analisadas. Tais dados nos levam a concluir que os textos legais são um registro segunda a perspectiva teórica por nós adotada, o que confirma nossa hipótese inicial. Com nosso trabalho esperamos ter contribuído na caracterização de um registro específico, aquele dos textos legais brasileiros. Estamos no momento completando o processamento dos dados do *LEX-BR-Ius* de maneira a poder realizar uma AMD completa no corpus e obter dados ainda mais refinados para uma descrição da linguagem utilizada nas normas legais do PB.

Agradecimentos

As autoras agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa (nº 88887.626989/2021-00) que permitiu esta investigação, à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelas bolsas de Iniciação Científica da equipe de compilação do corpus, e a Carlos Kauffmann pelo fornecimento dos dados do CVBR e etiquetagem do corpus.

Referências

- Barbera, M.; Onesti, C. (2009). Scheda Progetto di ricerca n. 9. Corpus Jus Jurium. In: Diadori, P (Ed.). *Progetto JURA: la formazione dei docenti di lingua e traduzione in ambito giuridico italo-tedesco*. Perugia: Guerra Edizioni, 2009. p. 349-351.
- Baker P., Hardie A. e McEnery T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh University Press.
- Berber Sardinha, T. (2010). A abordagem metodológica da Análise Multidimensional. *Gragoatá*, n. 29, p. 107-125.
- Berber Sardinha, T. (2013a). Variação entre registros da Internet. In: Shepherd, T. G.; Saliés, T. G. (Eds.). *Linguística da Internet*. Contexto, p. 55–85.
- Berber Sardinha, T. (2013b). *Pós-processador PT Tag Count*.
- Berber Sardinha, T., Kauffmann, C e Acunzo, C. M. (2014). Dimensions of register variation in Brazilian Portuguese. In: Pinto, M. V. (Eds.). *Multi-dimensional analysis: 25 years on a tribute to Douglas Biber*. John Benjamins Publishing Company.
- Biber, D. (1988). *Variations across speech and writing*. CUP.
- Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, v. 8, n. 4, Oxford: Oxford University Press, p. 243-257.
- Biber, D e Conrad, S. (2009). *Register, genre, and style*. CUP.
- Bick, E. (2014). PALAVRAS, a constraint grammar-based parsing system for Portuguese. In: T. Berber Sardinha, T.; Ferreira, T. São Bento (eds.), *Working with Portuguese corpora*, p. 279–302. Bloomsbury.
- Bick, E. (2000). *The Parsing System "Palavras"*: Automatic Grammatical Analysis of

- Portuguese in a Constraint Grammar Framework. Tese (PhD) – Aarhus University. Aarhus, Denmark: Aarhus University Press.
- Brasil (1990). *Lei nº 8.078, de 11 de setembro de 1990* (Código de Defesa do Consumidor).
- Brasil (1934). *Decreto nº 24.643 de 10 de julho de 1934* (Código de águas).
- Brasil (2010). *Lei nº 12.288, de 20 de julho de 2010* (Estatuto da Igualdade Racial).
- Carapinha, C (2018). A linguagem jurídica. Contributos para uma caracterização dos Códigos Legais. *REDIS: Revista de Estudos do Discurso*, n. 7. Disponível em: <https://ojs.letras.up.pt/index.php/re/article/view/6200>. Acesso em: 10 set. 2021.
- Coulthard, M.; Johnson, A.(eds.) (2010). *The Routledge handbook of forensic*. Routledge.
- Coulthard, M.; Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge.
- Ferrari, L. A.; Marques, C. G. de F. (2022) O *LEX-BR-Ius*: arquitetura e decisões na compilação de um corpus representativo das leis federais brasileiras. *ANTARES*, v.14, n.34. Disponível em: <http://www.uces.br/etc/revistas/index.php/antares/article/view/11150/5328>. Acesso em: 19 dez. 2022.
- Ferrari, L. A.; Marques, C. G. de F. (em preparação) O *LEX-BR-Ius*: apresentação do corpus e pesquisas preliminares.
- Goźdz-Roszkowski, S. (2012). Legal Language. In: Chapelle, Carol A. (Org.). *The Encyclopedia of Applied Linguistics*. John Wiley e Sons, p. 3281-3287.
- Lorz R. A. (2019). Creating Law with Language – Crossing Borders and Connecting Disciplines from the Perspective of Legislative Practice. In: Vogel F. (ed). *Legal Linguistics Beyond Borders: Language and Law in a World of Media, Globalisation and Social Conflict Relaunching the International Language and Law Association (ILLA)*. Duncker e Humblot GmbH, p. 5-8.
- Marques, C. G. de F. (2023). *Análise multidimensional dos textos legais federais brasileiros*. Dissertação (Mestrado em Estudos Linguísticos) – Universidade Federal de Minas Gerais, Belo Horizonte.
- McEnery, T. e Wilson A. (2001). *Corpus Linguistics*. Edinburgh UP, second edition.
- Tiersma, P. (1999). *Legal Language*. The University of Chicago Press.
- Tognini-Bonelli E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing Company.
- Sinclair, J. (2004). *Trust the Text: Language, Corpus, and Discourse*. Routledge.

Processamento da Linguagem Natural e Complexidade Sintática

Bruna R. da Silva¹

¹Programa de Pós-Graduação em Letras – Universidade Federal do Rio Grande do Sul (UFRGS)

Caixa Postal 91540-000 – Porto Alegre – RS – Brasil

thu_du@hotmail.com

Resumo: Este trabalho se insere nos estudos de Acessibilidade Textual e Terminológica. O *corpus* de estudo é a publicação digital *Aprendendo sobre vírus e vacinas*. Os *corpora* de contraste são: (1) conjunto de textos sobre cuidados paliativos, e (2) conjunto de sentenças dos Juizados Especiais Cíveis. O objetivo é mapear as características sintáticas desses *corpora*, a fim de verificar qual o *corpus* mais simples. Tais *corpora*, de características distintas, foram etiquetados pela gramática UDPipe e explorados conforme alguns critérios pré-estabelecidos de análise sintática. Os contrastes indicam que o *corpus* de estudo tem sintaxe mais simples, embora alguns dos pontos explorados mereçam especial atenção. Sendo assim, tais comparações vão enriquecer a análise e a discussão sobre a acessibilidade desse material, além de servir de base para o estudo como um todo.

Abstract: This work is part of the studies of Textual and terminological accessibility. The *corpus* of study is the digital publication *Learning about viruses and vaccines*. The contrasting *corpora* are: (1) set of texts about palliative care, and (2) set of judgments of the Special Civil Courts. The goal is to map the syntactic features of these *corpora*, in order to verify which *corpus* is the simplest. Such *corpora*, of distinct features, were tagged by the UDPipe Grammar and explored according to some pre-established syntactic analysis criteria. The contrasts indicate that the *corpus* of study has simpler syntax, although some of the points explored deserve special attention. Thus, such comparisons will enrich the analysis and discussion about the accessibility of this material, as well as serving as foundation to the study as a whole.

Introdução

Os estudos de Acessibilidade Textual e Terminológica (ATT) tratam dos padrões que tornam um texto mais ou menos complexo, sempre de acordo com o público para o qual esse texto é direcionado e produzido. Nesse sentido, a pesquisa de Doutorado no qual esse estudo está inserido busca a união da experiência docente com a pesquisa acadêmica, por meio da investigação da leitura e da compreensão de materiais, em tese, adaptados para um público com doze anos ou mais, por jovens e adolescentes do Ensino Fundamental II de escola pública de Porto Alegre-RS.

O objetivo principal da pesquisa é descrever e analisar se um livro da área da saúde, disponível *on-line*, adaptado para um público leitor jovem, é compreendido por esse público e de que forma. Inicialmente, o foco será a publicação digital *Aprendendo sobre vírus e vacinas*, da Editora da UFCSPA. Essa editora lançou várias publicações, todas na área da saúde, adaptadas para diferentes públicos. A única dessas obras direcionada para público jovem, com doze anos ou mais, foi escolhida para análise porque essa é a faixa etária que corresponde aos alunos da pesquisadora responsável, com os

quais será possível dar continuidade à pesquisa, num próximo momento, por meio de testes de compreensão leitora.

O recorte que se apresenta neste resumo faz parte de estudo piloto sobre o *corpus* de estudo, em que serão realizados contrastes do *corpus* de estudo com outros *corpora*. O objetivo é mapear as características dos *corpora*, a fim de verificar se de fato o *corpus* de estudo está acessível para o público ao qual se destina. Os *corpora* escolhidos para contraste foram dois: (a) conjunto de textos de divulgação científica, da área da saúde, sobre cuidados paliativos; e (b) conjunto de textos de sentenças dos tribunais de pequenas causas, da área jurídica.

Ao lermos os três materiais, a hipótese principal é que o *corpus* de estudo seja o mais simples. A segunda hipótese é que, ainda que o *corpus* de estudo seja o mais simples entre os três, precisa de mais adaptação e simplificação para de fato se tornar acessível e compreensível para o público de alunos do Ensino Fundamental II.

A fim de verificar nossas hipóteses, todos os *corpora* foram submetidos ao mesmo procedimento, com as mesmas ferramentas computacionais: anotação pela UDPipe e buscas pelo Interrogatório. Ambas as ferramentas serão contextualizadas e explicadas na seção de Métodos.

Os resultados validaram a aposta inicial, de que o *corpus* de estudo seria o mais simples, de acordo com os critérios pesquisados. Porém, a segunda hipótese não foi confirmada, pois não houve critério em que o *corpus* de estudo fosse considerado o mais complexo, necessitando de atenção e simplificação.

Referencial Teórico

De maneira geral, a palavra “acessibilidade” corresponde a “ser acessível”. Então, acostumamo-nos a relacionar essa palavra com a inclusão, com os espaços, com os recursos. Mas, tal acessibilidade também pode ser analisada do ponto de vista textual, das palavras, das frases, dos parágrafos, dos textos, bem como da informação, da compreensão. É dessa forma que enxergamos a acessibilidade neste trabalho, alinhada ao que mencionam Finatto e Motta (2019, p. 317):

Hoje, porém, se entende que a noção de *acessibilidade* não é algo apenas associado a espaços físicos. Afinal, para as pessoas que buscam alguma inserção em diferentes contextos, sejam esses físicos, sejam abstratos, pode haver barreiras de vários tipos. É o caso, por exemplo, de muitos materiais escritos que, por várias razões, se tornam “ambientes” totalmente inacessíveis a um grande número de pessoas, portadoras de necessidades especiais ou não.

As autoras Fulgêncio e Liberato (1992) já discutiam, nos anos 1990, a importância da adequação da leitura ao público-alvo ao qual se destina. Segundo Perini (1988), que acreditava que só se aprende a ler, lendo, “[...] a leitura funcional nascerá do convívio com o material escrito adequado, e somente dele” (PERINI, 1988, p. 81). Dessa forma, o autor propõe que textos, quanto ao nível de dificuldade de leitura, devem ser graduados a diferentes públicos, “[...] de modo que um texto da terceira série fosse significativamente mais simples do que um de oitava série, ou de nível universitário” (PERINI, 1988, p. 82).

Baseados nessas referências, o interesse deste estudo está no texto do livro *Aprendendo sobre Vírus e Vacinas*, que, conforme os autores, foi adaptado para jovens com idade entre doze anos ou mais, alunos de Ensino Fundamental II. Mas, como a

publicação não foi testada de fato, na prática, não há como saber se o público a compreende. Em função disso, o estudo pretende explorar esse material, contrastando-os com outros *corpora* e realizando testes de compreensão leitora com esse público.

Assim, nesse momento, analisaremos a sintaxe do texto, em comparação com outros dois *corpora* também escritos para pessoas leigas. Como auxílio para tal análise, utilizamos o NILC-Metrix, um sistema computacional composto por 200 métricas propostas em estudos sobre discurso, psicolinguística, linguística cognitiva e computacional, para avaliar a complexidade textual do português brasileiro (PB). Esse sistema é a adaptação mais moderna da ferramenta Coh-Metrix para o inglês.

Desenvolvida por pesquisadores da Universidade de Memphis, a ferramenta Coh-Metrix é uma ferramenta de análise de textos que “[...] calcula índices que avaliam a coesão, a coerência e a dificuldade de compreensão de um texto (em inglês) [...]” (SCARTON, ALUÍSIO, 2010, p. 47). No Brasil, a ferramenta foi adaptada para o português brasileiro por pesquisadores do Núcleo Interinstitucional de Linguística Computacional da USP e da UFSCar (NILC-USP), ao longo dos últimos treze anos.

No índice do NILC-Metrix, é possível selecionar métricas específicas de complexidade sintática. Entre essas métricas estão três da utilizadas neste estudo: quantidade de conjunções subordinativas em relação ao total de conjunções, quantidade de orações subordinadas em relação ao total de orações, e quantidade de orações com voz passiva em relação ao total de orações.

Outro estudo que também mediu a complexidade sintática de textos literários para público neoleitor (adultos que estão sendo alfabetizados tardiamente) foi o de Rodrigues, Freitas e Quental (2013). As autoras investigaram a Coleção Literatura para Todos 1, de 2006. Em função do público pretendido, imaginava-se que a escrita dessas obras fosse mais acessível linguisticamente. Foi em função disso que as autoras investigaram a complexidade sintática dessas obras de maneira objetiva e quantitativa, por meio de ferramentas computacionais. As ferramentas utilizadas na época foram o analisador morfossintático PALAVRAS (BICK, 2000) e o programa Coh-Metrix Port (ALMEIDA e ALUÍSIO, 2009). As buscas utilizadas pelas autoras naquele momento, que foram reutilizadas em nossa pesquisa atual, foram: quantidade de verbos por sentença, quantidade de vírgulas por sentença, quantidade de elementos intercalados por sentença e quantidade de orações reduzidas de gerúndio em relação ao total de orações.

Neste momento, em que novas e diferentes ferramentas de Linguística Computacional estão a nossa disposição, utilizamos a UDPipe (STARAKA *et al.*, 2016) para etiquetar os *corpora* e o Interrogatório (de SOUZA e FREITAS, 2019) para realizar as buscas nesses *corpora* etiquetados.

Materiais

Os três *corpora* utilizados nesse estudo foram: (a) *corpus* de estudo: livro Vírus e Vacinas, (b) *corpus* de contraste I: conjunto de textos de divulgação científica sobre cuidados paliativos, e (c) *corpus* de contraste II: conjunto de sentenças dos tribunais de pequenas causas.

Um *corpus* de estudo é a base de uma pesquisa, é o material, o compilado de textos que pretendemos descrever e avaliar. Já o *corpus* de contraste é aquele compilado de

textos utilizado na comparação, a fim de aproximar ou distanciar os materiais quanto aos mais diferentes aspectos: semânticos, sintáticos, léxicos, estruturais, quantitativos, etc.

É importante ressaltar que esses materiais foram escolhidos porque se destinam à população em geral. Dessa forma, todos deveriam ser de leitura e entendimento acessíveis. Enquanto o *corpus* de estudo tem como público alunos do Ensino Fundamental II, o *corpus* de contraste I é formado por textos sobre cuidados paliativos para pessoas leigas. Pessoas leigas são aquelas que não têm conhecimento sobre o assunto, geralmente cuidadores e familiares de pessoas que necessitam de (ou fazem escolha por) cuidados paliativos, então, os textos precisam ser compreendidos por elas. Da mesma forma, o *corpus* de contraste II é formado por sentenças de tribunais de pequenas causas, ou seja, englobam processos de menor complexidade, que não ultrapassem 40 salários mínimos e qualquer pessoa pode utilizar sem a necessidade de um advogado, sendo assim, tais sentenças também precisam ser acessíveis à população.

O *corpus* de estudo é o livro *Aprendendo sobre Vírus e Vacinas*, da Editora da UFCSPA, publicado no ano de 2020. O livro tem 72 páginas e está disponível para *download* gratuito no site da Universidade. O projeto de extensão de mesmo nome, *Aprendendo sobre Vírus e Vacinas*, objetiva auxiliar o aprendizado dos alunos dos anos finais do Ensino Fundamental no aprendizado sobre vírus e vacinas utilizadas na sua prevenção (RODRIGUES, 2020, p. 65). Os autores são estudantes de Biomedicina na UFCSPA que, juntamente com o professor Luiz Carlos Rodrigues Jr., diante da pandemia do novo Coronavírus, criaram histórias, baseadas nas suas pesquisas e estudos, adaptando-as para diferentes públicos. O único direcionado para a idade dos alunos (12+) com os quais será possível dar continuidade à pesquisa num segundo momento, por meio de testes de compreensão leitora, é o livro *Aprendendo sobre Vírus e Vacinas*, que, conforme os autores:

“Apresenta um panorama geral das principais epidemias virais, incluindo a do SARS-CoV-2, aspectos da estrutura, da infecção e da replicação viral no hospedeiro, a resposta imunológica formada e o mecanismo imunogênico das vacinas. Cada parte é ilustrada com imagens simplificadas e lúdicas dos sistemas e dos mecanismos envolvidos.” (RODRIGUES, 2020, p. 65).

Os *corpora* de contraste foram escolhidos porque também lidam com um público específico: a população em geral.

O *corpus* de contraste I também é da área da saúde, foi organizado durante a Dissertação de Esteves (2023), e é composto por conjuntos de textos de divulgação científica sobre cuidados paliativos. A autora explica que, a partir de busca no Google, foram selecionados textos com as seguintes palavras-chave: a) “cuidados paliativos” e b) “cuidados paliativos” + “Ministério da Saúde”. Essa mesma pesquisa foi realizada em três datas diferentes: 1) 12 de maio de 2021, com a seleção de 10 textos; 2) 15 de junho de 2021, com a seleção de 5 textos; e 3) 19 de setembro de 2021, com a seleção de 15 textos. Ao selecionar os textos, foram usados alguns critérios de inclusão: a) apenas em português brasileiro; b) de acesso livre e gratuito; c) de fontes idôneas, como instituições do Governo, portais de notícias reconhecidos e entidades bem estabelecidas; d) preferência a textos escritos originalmente em português; e) preferência a textos jornalísticos ou de divulgação científica. Em função da pandemia, foram compilados outros 15 textos, com resultados para buscas Google com as palavras-chave: “cuidados paliativos” + “CPI da pandemia”. Essa última seleção ocorreu em dois momentos: 1) 12

de novembro de 2021 e 2) 4 de fevereiro de 2022. Foram utilizados os mesmos critérios de inclusão e exclusão das compilações anteriores, totalizando um corpus de 45 textos sobre cuidados paliativos. (ESTEVES, 2023)

O *corpus* de contraste II é da área jurídica, foi organizado na Tese de Motta (2022), e compreende um total de 440 Sentenças Judiciais dos Juizados Especiais Cíveis (JECs) do Poder Judiciário do Rio Grande do Sul (PJRS). Essas Sentenças representam as decisões tomadas pelos juízes em relação aos pedidos das pessoas que procuram os JECs para resolver algum conflito. O Juizado Especial Cível tem como atribuição a conciliação, o processo e o julgamento de ações cíveis de menor complexidade, cujo valor de causa não ultrapasse 40 salários mínimos. As Sentenças foram coletadas do *site* do Tribunal de Justiça do Rio Grande do Sul, entre os anos de 2018 e 2019. As temáticas das Sentenças são: água, cadastro de inadimplentes, energia elétrica, plano de saúde, prestação de serviços e telefonia (de forma que o *corpus* ficasse variado). O *corpus* organizado compreende um total de 440 sentenças. (MOTTA, 2022)

Métodos

Primeiramente, os três *corpora* foram formatados em arquivo de texto, seguindo o padrão UTF8, necessário para anotação em UDPipe. A ferramenta UDPipe, de código aberto, “[...] realiza sequencialmente as etapas de tokenização (segmentação do texto em unidades básicas, como palavras e sinais de pontuação), anotação gramatical, lematização e análise de dependências em qualquer *corpus* que esteja no formato CoNLL-U.” (de SOUZA e FREITAS, 2021, p.1043). Utilizamos esse modelo de anotação por que:

O modelo fornecido para o português (versão 2.5) tem índices de acerto (F1) de 96.4%, 95%, 87.2% e 83.1% para os níveis de classes gramaticais (POS), características morfológicas (feats), dependência sintática (unlabeled attachment score (UAS) e relação de dependência sintática (labeled attachment score (LAS)), respectivamente. .” (de SOUZA e FREITAS, 2021, p.1043)

Esses *corpora* anotados foram processados no Interrogatório, uma ferramenta para busca e revisão de *corpora* anotados (de SOUZA e FREITAS, 2019). Essa ferramenta realiza buscas sintáticas em arquivos anotados em formato CoNLL-U.

É importante mencionar que o *corpus* de sentenças judiciais, por ser muito extenso, teve de ser dividido em duas partes, pois as ferramentas não conseguiram realizar seu processamento de uma única vez. Então, as buscas e resultados para esse *corpus* em específico foram dobradas. Todos os resultados foram somados para obtermos um resultado final do todo do *corpus*.

A partir disso e de estudos anteriores já mencionados nas referências, foi verificada a quantidade de ocorrências dos seguintes indícios de complexidade textual:

- Quantidade de verbos por período;
- Quantidade de vírgulas por período;
- Quantidade de elementos explicativos intercalados por período;
- Orações subordinadas em relação às orações;
- Voz passiva em relação às orações;
- Orações reduzidas de gerúndio em relação às orações;
- Conjunções subordinativas em relação a todas as conjunções.

Essas buscas específicas aconteceram por terem sido delimitadas como indícios de complexidade textual, conforme os estudos de Rodrigues, Freitas e Quental (2013) e conforme apontamentos do NILC-Metrix (LEAL *et al*, 2022).

Algo que notamos desde o início é que as buscas a partir dessa anotação nem sempre são óbvias e/ou diretas. Por exemplo, para analisarmos o primeiro item, a quantidade de verbos por período, não somente buscamos os verbos, pois há verbos anotados de outra maneira, como auxiliares, por exemplo, ainda que sejam verbos principais, dependendo da frase. Alguns exemplos estão na Tabela 1 a seguir:

Tabela 1: Exemplos de verbos auxiliares como principais

<i>Corpus</i>	Exemplo
Vírus e Vacinas	Os neutrófilos são os primeiros a chegar.
Cuidados Paliativos	O que leva um médico a atuar com quem está na reta final da vida?
Sentenças Jurídicas	A parte Demandada é prestadora de serviço de energia elétrica, logo fornecedora .

Então, para encontramos a quantidade total de verbos, foram necessárias duas buscas separadas, nas quais a quantidade de verbos encontrada em cada *corpus* e em cada busca foi somada. Primeiro, buscamos os verbos. Em seguida, os verbos auxiliares com função de verbo principal, que buscou verbos como os da Tabela 1.

Para encontrarmos as vírgulas, buscamos pela pontuação do texto, todas que eram iguais à vírgula. Mas, para buscarmos os elementos intercalados também não foi possível uma única busca, pois não há como a ferramenta extrair todos de uma única vez. Assim, para encontrarmos esses elementos intercalados, foram realizadas cinco buscas diferentes para cada *corpus*: aposto intercalado, adjunto adverbial anteposto ao verbo, oração adverbial anteposta ao verbo, oração adjetiva explicativa e oração adjetiva reduzida. Na Tabela 2 estão alguns exemplos de elementos intercalados:

Tabela 2: Exemplos de elementos intercalados

<i>Corpus</i>	Exemplo
Vírus e Vacinas	PARA ESCREVER ESTE LIVRO , NÓS TAMBÉM ESTUDAMOS MUITO! (oração adverbial)
Cuidados Paliativos	E, na Bélgica , ocorre, atualmente, a mesma situação jurídica. (adjunto adverbial)
Sentenças Jurídicas	A parte autora alega perda de carne bovina , entre outros alimentos , que se encontravam congelados para consumo próprio, em dois frízers, no montante de R\$1.690,00 (um mil seiscentos e noventa reais). (oração adjetiva explicativa)

Ao final das buscas, todos os números encontrados em cada *corpus* foram somados e relacionados com a quantidade de períodos. Dessa forma, encontramos a média de elementos intercalados por período e por *corpus*.

Para buscarmos as orações subordinadas, foi preciso realizar várias buscas diferentes para cada *corpus*. Orações subordinadas substantivas subjetivas e objetivas diretas, além de casos de coordenação entre elas; orações subordinadas adjetivas e adverbiais, além de casos de coordenação entre elas. Todos os números foram somados em cada *corpus*.

Para encontrarmos a voz passiva também foi mais tranquilo, pois foi possível utilizar apenas um comando por *corpus*. Entretanto, para encontrarmos as orações reduzidas de gerúndio, também foram necessárias algumas buscas para cada *corpus*: orações subordinadas adjetivas reduzidas de gerúndio, orações subordinadas adverbiais reduzidas de gerúndio, orações subordinadas substantivas que podem ser reduzidas de gerúndio.

E, por fim, para encontrarmos as conjunções subordinativas, foi possível somente uma busca, porém, para relacionarmos com a quantidade total de conjunções, foram necessárias duas buscas (conjunções subordinativas + conjunções coordenativas).

Resultados

Primeiramente, os dados gerais dos *corpora* são apresentados na Tabela 3 a seguir.

Tabela 3: Informações sobre os *corpora*

	<i>Corpus</i> de estudo	<i>Corpus</i> de Contraste (1) - SAÚDE	<i>Corpus</i> de Contraste (2) - JUSTIÇA
<i>Tokens</i>	4.465	33.0731	702.511
Sentenças	342	13.883	30.604
Quantidade de documentos	1 publicação digital	45 textos do Google	440 Sentenças Judiciais

É importante ressaltar que, embora os *corpora* tenham extensões diferentes, o que é visível pelos dados da Tabela 3, isso não compromete os resultados e as comparações, uma vez que todas as análises foram relacionadas de acordo com os dados de cada *corpus*. Por exemplo, a quantidade de verbos, de vírgulas e de elementos intercalados do *corpus* de estudo foi relacionado com a quantidade de sentenças desse *corpus*, 342. Já a quantidade de verbos, de vírgulas e de elementos intercalados do *corpus* de contraste (2) foi relacionada com a quantidade de sentenças do *corpus* da justiça: 30.604. Dessa forma, os resultados foram sempre relativos ao próprio *corpus*, sua extensão, quantidade de sentenças e características, o que permitiu as comparações a seguir.

A Tabela 4 a seguir mostra os resultados finais, de cada soma de critérios, obtidos com todas as buscas, separadas por *corpus* e por critério pesquisado:

Tabela 4: Resultados Gerais

	<i>Corpus</i> de estudo	<i>Corpus</i> de Contraste (1) - SAÚDE	<i>Corpus</i> de Contraste (2) - JUSTIÇA
Verbos/Período	1,38	2,47	2,24
Vírgulas/Período	0,37	1,3	1,5
Elementos intercalados/Período	0,23	0,54	0,66
Subordinadas/Orações	0,39	0,49	0,49
Voz Passiva/Orações	0,04	0,06	0,05
Reduzidas de Gerúndio/Orações	0,01	0,03	0,05
Conjunções Subordinativas/Conjunções	0,39	0,03	0,58

Os números grifados em negrito apontam para maior resultado em cinco critérios para o *corpus* de sentenças jurídicas. Em relação às orações subordinadas, houve empate de resultado entre os dois *corpora* de contraste. Ainda assim, o *corpus* de cuidados paliativos é o segundo mais complexo, com resultados maiores em três critérios.

Conforme a nossa hipótese inicial, o *corpus* de estudo, sobre vírus e vacinas, é o menos complexo em relação aos outros dois, pois não apresenta resultados maiores em nenhum dos critérios pesquisados. Apesar disso, nestes critérios pesquisados, a segunda hipótese, de que o *corpus* de estudo ainda poderia ser adaptado e simplificado para o público ao qual se destina, não se confirmou, visto que não houve resultado maior em nenhum quesito, então, não há como afirmar a segunda hipótese.

Considerações

O objetivo principal do estudo foi mapear as características sintáticas dos *corpora* e verificar qual deles seria o mais simples, sintaticamente. O objetivo foi realizado e o resultado principal comprovou a hipótese de que o *corpus* de estudo é o mais simples, sintaticamente, entre os *corpora* analisados e de acordo com os critérios estabelecidos, baseados em estudos anteriores.

Porém, algumas dificuldades surgiram ao longo do caminho. A primeira delas foi a familiarização com os termos e a anotação realizada. Não é tão óbvio encontrar as estruturas da gramática tradicional, pois a anotação na gramática UDPipe nem sempre segue a mesma lógica, então, a busca precisa ser ajustada (e compreender a anotação exige abstração, foco e experimentação).

No passo seguinte, as buscas no Interrogatório, foi preciso também experimentação nas buscas, pois nem sempre o que pensamos de acordo com a gramática tradicional já internalizada em nosso raciocínio, é o que a busca realizada nos traz como resultado. Às vezes, os caminhos precisam ser refinados, acrescentando mais de uma condição, e/ou filtros. Então, em vários dos critérios foi preciso mais de uma busca, sobretudo nos elementos intercalados, critério mais complicado, em nosso entendimento e experiência.

Além disso, o *corpus* de sentenças precisou ser dividido em duas partes, pois a ferramenta não funcionou com ele por completo. Isso tornou as buscas nesse *corpus* especificamente mais trabalhosas, pois foram o dobro de pesquisas e de somas.

Ainda assim, acreditamos que o estudo contribuiu muito para a familiarização com os termos, os critérios e os funcionamentos desses sistemas computacionais, os quais não conhecíamos, nem nunca havíamos experimentado. Além disso, é somente na prática que aprendemos a lidar com algo que não é do nosso uso frequente e familiar. Adaptar nosso pensamento e lógica tão acostumados com a gramática tradicional também foi um exercício importante.

Por fim, de forma geral, acreditamos que o estudo foi útil não só para o aprendizado relacionado às ferramentas e à análise linguística, como também para as discussões sobre complexidade textual de diversos *corpora* que podem ser explorados dessa mesma maneira. Além disso, acreditamos que tais contrastes também vão enriquecer a análise e a discussão sobre a acessibilidade do *corpus* de estudo para esse público, servindo de base para as etapas subsequentes da pesquisa e da Tese.

Referências

- de SOUZA, E.; FREITAS, C. ET: uma Estação de Trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente. In: **WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA** (TILic 2019), VI., 2019. Salvador. Proceedings [...]. Salvador: SBC, 2019. p. 15-18.
- de SOUZA, E.;FREITAS, C. Sujeito oculto às claras: uma abordagem descritivo-computacional. In: **Revista Estudos da Linguagem**. Belo Horizonte, v. 29, n. 2, p. 1033-1058, 2021.
- ESTEVES, F. F. **DEFINIÇÕES ACESSÍVEIS: POR UMA LINGUAGEM SIMPLES EM CUIDADOS PALIATIVOS**. 157f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.
- FINATTO, M. J. B. ; MOTTA, E. Terminologia e Acessibilidade: novas demandas e frentes de pesquisa. **REVISTA GTLEX**, v. 2, p. 316-356, 2019.
- FULGÊNCIO, L.; LIBERATO, Y. **Como Facilitar a Leitura**. São Paulo: Contexto, 1992.
- LEAL, Sidney Evaldo, DURAN, Magali Sanches, SCARTON, Carolina Evaristo, HARTMANN, Nathan Siegle, ALUÍSIO, Sandra Maria. **NILC-Metrix: avaliando a complexidade da linguagem escrita e falada no português brasileiro**. CoRR abs/2201.03445 (2022). Disponível em: <https://arxiv.org/abs/2201.03445> Acesso em: 11 jun. 2023.
- MOTTA, E. **SENTENÇAS JUDICIAIS E LINGUAGEM SIMPLES: um encontro possível e necessário**. 411f. Tese (Doutorado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2022.
- PERINI, M. A. A Leitura Funcional e a Dupla Função do Texto Didático. In: ZILBERMAN, Regina; SILVA, Ezequiel Theodoro (org.). **Leitura: Perspectivas Interdisciplinares**. São Paulo: Ática, 1988.
- RODRIGUES, E. C.; FREITAS, C.; QUENTAL, V. Análise de inteligibilidade textual por meio de ferramentas de processamento automático do português: avaliação da Coleção Literatura para Todos. In: **Letras de Hoje**, Porto Alegre, v.48, n. 1, p.91-99, jan./mar. 2013.
- RODRIGUES, L. C. J. **Aprendendo sobre vírus e vacinas** [recurso eletrônico]. Porto Alegre: Editora da UFCSPA, 2020. ISBN 978-65-87950-28-0. Disponível em: <https://www.ufcspa.edu.br/vida-no-campus/editora-da-ufcspa/obras-publicadas>. Acesso em: 30 set. 2021.
- SCARTON, C.; ALUISIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o português. **LinguaMática 2** (2010): p. 45-62.
- STRAKA, M.; HAJIC, J.; STRAKOVÁ, J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: **TENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION** (LREC'16), 10th., Portorož, Proceedings [...]. Portorož: ELRA, 2016. p. 4290-4297.

Formal features for a syntactic-semantic classification of predicative adjectives in Brazilian Portuguese

Ryan Marçal Saldanha Magaña Martinez¹, Oto Araújo Vale¹

¹Programa de Pós-graduação em Linguística – Universidade Federal de São Carlos (UFSCar) – São Carlos – SP – Brazil

ryan.martinez@estudante.ufscar.br, otovale@ufscar.br

Abstract. *Syntactic-semantic classification of words relies ultimately on the selection of appropriate formal features. This is a bibliographic study discussing the applicability and relevance to Brazilian Portuguese of features proposed to correlate to semantics in large-scale work on predicative adjectives. These include distributional features of subjects, complements, prepositions, and copular verbs, as well as transformational features (i.e., those obtained by information-preserving rephrasing of sentences). The results of this bibliographic search suggest that 40 of the formal properties registered in other work are promising for dividing Brazilian Portuguese adjectives into syntactic-semantic classes.*

1. Introduction

Large-scale efforts for a syntactic-semantic classification of Brazilian Portuguese (BP) predicates focus mainly on verbs [Cançado et al. 2018] and nouns [Barros 2014], [Santos 2015], [Calcia 2022], [Rassi 2023]. Predicative adjectives (A_{pred}), nonetheless, are so far an underexplored phenomenon when it comes to large-scale classification of BP predicates.

Such adjectives select the arguments in sentences with copular verbs. A transformational property – i.e., the acceptability of an information-preserving rephrasing – characterizes A_{pred} : its use in adnominal function should be equivalent to a relative clause consisting of a copular verb plus the adjective [Vendler 1968]: *um aluno ansioso para voltar pra casa* (a student anxious to go home) = *um aluno que está ansioso para voltar pra casa* (a student who is anxious to go home).

Large-scale studies on the syntax-semantics of A_{pred} depend on the identification of relevant formal properties. Since natural language sentences constitute an infinite set, with an ever-growing lexicon [Harris 1991], the number of properties a predicate might show cannot be grasped in its entirety. Additionally, not all of them are correlated to semantics. Different and complementary approaches can be used to define relevant features. Bibliographic research is the most fundamental of such approaches, saving the researcher from spending a long time with corpus research and introspection only to rediscover correlations. Systematic descriptions of European Portuguese and other languages are available and their findings might prove useful for identifying such features.

The objective of this paper is thus to provide a discussion of formal properties employed in large-scale work for the classification of A_{pred} . This is expected to point out directions for large-scale studies of BP A_{pred} .

2. The studies

The discussion on formal features of adjectives and their correlation with semantic features is too extensive to be compiled in its entirety. Also, part of these studies are in languages we have no access to. To make this study possible, we have included exclusively works in Portuguese, French, or English that classify over a thousand A_{pred} , i.e., large-scale investigations into this type of predicate, which we believe to make assertions with strong empirical support. Although this leaves aside theoretically important accounts of adjectives, such as [Vendler 1968], such texts are contemplated indirectly, since their key ideas are applied to the empirical investigations described in our sources.

The selection led to a body of six studies. All of them follow the Lexicon-Grammar theoretical and methodological framework, which proposes the registration of properties of natural language predictors in matrices [Gross 1975] [Gross 1981]. One of them [Valetopoulos 2003] combined Lexicon-Grammar with the *classes d'objets* framework [G. Gross 1994], which aims at registering the semantic types of arguments. A qualitative and quantitative synthesis of these studies is provided in Table 1.

Table 1. Languages, objects, lexemes, features, and classes per large-scale study of predicative adjectives

Reference	Target language(s)	Object	Number of lexemes	Number of features	Number of classes
Picabia 1978	French	All A_{pred}	“Around 4,000”	86	16
Casteleiro 1981	European Portuguese	A_{pred} with complement clauses	2,039	47	11 (+3 subclasses)
Nam 1996	Korean	All A_{pred}	6,500	unspecified	15
Valetopoulos 2003	Greek and French	All A_{pred}	unspecified (around 7,000 estimated)	unspecified	unspecified
Carvalho 2007	European Portuguese	Intransitive A_{pred} with human subjects	4,250	unspecified	4 (+12 subclasses)
Messina 2019	Italian	A_{pred} with complement clauses	1,732	unspecified	10

We have searched each of these six studies for discussions on formal properties having semantic counterparts, verifying through introspection and corpora whether their application to BP is of relevance. As expected in bibliographic research involving multiple languages, all proposals discussed constructions or syntactic-semantic correlations which are not observable in BP. These are omitted from this paper due to

space restrictions. The results are presented in two sections: one discussing distributional features and the other dealing with transformational features.

3. Distributional features

In Lexicon-Grammar, distributional features are those of base sentences, i.e., sentences containing a predicate, its subject (N_0), complements (N_1 , N_2), and obligatory grammatical elements, such as copular verbs (V_{cop}) and prepositions (Prep), in an unmarked order [Harris, 1968], [Harris, 1991]. In Portuguese, the base sentences of A_{pred} can be characterized as follows [Casteleiro 1981], [Carvalho 2007], with optional elements between parentheses: $N_0 V_{cop} A_{pred} (Prep N_1)$.

In BP, both N_0 and N_1 may be human (N_{hum}), non-human (N_{-hum}), finite clauses (*Que F*), infinitive clauses (V_{inf}), factive clauses (*o fato de que F* or *o fato de V_{inf}*) or, more rarely, conditional clauses (*se F*). Both finite and factive finite clauses can be either in the indicative (F_{ind}) or subjunctive (F_{sub}) mood. Arguments of the type N_{hum} are usually defined as those accepting substitution with given names or certain interrogative pronouns which translate to English “who” (in BP, *quem*); N_{-hum} covers arguments which do not pass these tests. Since N_{hum} and N_{-hum} are defined through formal criteria, they count as formal properties.

Psychological predicates are sure to accept N_{hum} (not exclusively) at least in one type of sentence, since psychological states are by definition human manifestations [Gross 1975], [Casteleiro 1981]. This applies to A_{pred} such as *alegre* (happy), *assustado* (scared), and *cansado* (tired). The same is true of A_{pred} denoting human behavior [Nam 1996], such as *inteligente* (clever) and *corajoso* (brave), and A_{pred} denoting nationalities, philosophical positions, and diseases [Carvalho 2007].

Alternation between N_{hum} and clauses in subject position has been associated with human behavior predicates [Nam 1996]: both *Maria é assustadora* (Maria is scary) and *é assustador entrar numa caverna* (it is scary to go into a cave) are acceptable sentences, but the latter is unacceptable for predicates taking an experiencer subject: **é assustado entrar numa caverna* (*it is scared to go into a cave). Behavior A_{pred} are said to only accept subject clauses if they are also evaluative A_{pred} [Valetopoulos 2003]. Thus, *é horrível que ele se comporte assim* (it is horrible that he behaves like this) is acceptable, but **é severo que ele se comporte assim* (*it is severe that he behaves like this) is not. Acceptability of the latter can be increased by adding *da parte de* (“of” as in “it is bold of you to assume this”). Relatedly, intransitive A_{pred} with subject clauses have also been associated to evaluation or appreciation by the speaker [Messina 2019], such as in *é certo fazer críticas construtivas* (it is right to make constructive criticism).

Adjectives denoting extrinsic properties of human beings, including “states of mind”, usually accept both finite and infinitive object clauses, which distinguishes them from other A_{pred} with human subjects [Messina 2019]. This can be exemplified by the fact that both *estou ansioso para o meu time ganhar* (I am anxious for my team to win) and *estou ansioso para que meu time ganhe* (I am anxious for “that my team wins”) are acceptable sentences, whereas **estou capacitado para que ganhe* (*I am skilled that I win) sounds odd in comparison with *estou capacitado para ganhar* (I am skilled to win).

In turn, adjectives accepting non-propositional N_{-hum} subjects (not exclusively) have been associated with predicates describing information from the five senses:

vision, hearing, taste, smell, and touch [Nam 1996]. In BP, these are exemplified by, respectively, *quadrado* (square), *grave* (low-pitched), *doce* (sweet), *cheiroso* (scented), *áspero* (rough). It is not uncommon for these adjectives to accept human or propositional subjects also in a metaphorical sense: *doce* might mean “pleasant” when it extends to propositions, e.g., *é doce andar no parque* (walking through the park is sweet), and “gentle” when it applies to human beings: *Carlos é um rapaz muito doce* (Carlos is a very sweet young man).

Factive clauses have been associated with predicates imprinting a semantically factive reading on its subordinate clauses [Gross 1975], [Picabia 1978], [Casteleiro 1981]: *o fato de João ter ido embora é comprometedor* (the fact that João is gone is compromising) entails *João foi embora* (João is gone). There seems to be a subclass of intransitive A_{pred} denoting different degrees of truth values [Picabia 1978], [Messina 2019], such as *verdadeiro* (true), *provável* (likely), and *falso* (false), which accepts only non-factive subject clauses. Relatedly, in finite complement clauses there is a tendency for the subjunctive mood to be used with non-factive A_{pred} denoting uncertainty or subjectivity, whereas the indicative is related to factual propositions [Casteleiro 1981]: *é bom que o João compra arroz* (it is good that João buys rice) entails that João buys rice, whereas *é bom que o João compre arroz* (it is good that João buy rice) does not. Finally, the small set of A_{pred} introducing their clauses with *se* (if/whether), such as *incerto* (uncertain) or *duvidoso* (doubtful), was suggested to convey doubt [Casteleiro 1981], [Messina 2019].

Acceptability of both subject and object clauses is associated to A_{pred} denoting equivalence, which are said to be symmetrical, i.e., inverting subject and complement conveys the same information [Picabia 1978], [Casteleiro 1981], [Nam 1996], [Messina 2019]. In BP, this behavior can be observed in A_{pred} such as *equivalente a* (equivalent to) and *parecido com* (similar to): *andar de moto é parecido com andar de bicicleta* (riding a motorcycle is similar to riding a bike) conveys the same information as *andar de bicicleta é parecido com andar de moto* (riding a bike is similar to riding a motorcycle).

Physical and psychological human state A_{pred} can be distinguished by the acceptance of non-coreferential causative clauses [Valetopoulos 2003]. In BP these are introduced with preposition *de*, as in *Pedro está cansado de discutir esse assunto* (Pedro is tired of discussing this issue). **Pedro está cansado de a Maria discutir esse assunto* (**Pedro is tired of Mary discussing this issue*) is marginal at best, but both *Pedro está triste de descobrir isso* (Pedro is sad to find that out) and *Pedro está triste de a Maria ter descoberto isso* (Pedro is sad for Maria to find that out) are acceptable.

Portuguese has two copular verbs: *ser* and *estar*. Although these are commonly assumed to relate, respectively, to inherent and accidental adjectives, it has been shown that this is context-dependent [Casteleiro 1981], [Carvalho 2007]. Still, A_{pred} with object clauses participating in volitive constructions (such as the imperative) mostly accept *ser*, whereas those which do not accept these constructions form sentences with *estar*, e.g., *seja hostil!* (be hostile!) is quite acceptable, but *?esteja ansioso!* (?be anxious!)¹ has much lower acceptability [Casteleiro 1981]. Also, nationality and belief adjectives

¹ We use question marks before sentences to indicate low, but non-zero, acceptability.

accept *ser*. This verb generally combines with congenital or chronic diseases, whereas *estar* occurs with temporary ones [Carvalho 2007].

A subclass of A_{pred} which selects exclusively locative complements has been proposed [Picabia 1978]. These are all stative, with no assignment of source, path, or destination semantic roles [Nam, 1996]. In BP, they correspond to A_{pred} such as *próximo a* (next to), *paralelo a* (parallel to), and *exterior a* (external to). Prepositions *a* and *em* seem to be common choices with these adjectives, but others, such as *de* (*perto de* “near to”), are not impossible. Adjectives denoting orientation [Nam 1996] will usually take *para* N_{hum} : *a torre está (inclinada + voltada) para um prédio* (the tower is (inclined to + facing) a building).

When preposition *para* introduces N_{hum} , these are commonly experiencers. The same adjectives which accept this *para* also accept *da parte de* (roughly “of” as in “it is bold of you to assume this”) introducing stimuli [Casteleiro 1981]. Acceptance of *da parte de* has been associated also to behavior A_{pred} [Messina 2019]. This alternation can be shown, e.g., in *interessante* (interesting): *apresentar esse seminário foi interessante para a Maria* (presenting this seminar was interesting for Maria) means that Maria found the experience of presenting the seminar interesting; *apresentar esse seminário foi interessante da parte da Maria* (it was interesting of Maria to present this seminar) means that the speaker found the stimulus “Maria presented this seminar” interesting.

Korean has a different type of alternation between experiencer and stimulus using postpositions, in which the stimulus is not necessarily a participant in a complement clause [Nam 1996]. Certain adjectives, such as *indiferente* (indifferent) have similar alternations in BP which might be worth testing: *Zé é indiferente a viajar de avião ou de ônibus* (Zé is indifferent to travelling by plane or by bus) is more or less equivalent to *viajar de avião ou de ônibus é indiferente para o Zé* (travelling by plane or by bus is indifferent to Zé). A large-scale study is required to pinpoint how widespread this alternation is in BP.

When *para* introduces complement clauses rather than human or non-human nouns, it usually denotes a purpose for N_0 [Casteleiro 1981], e.g., *João é (importante + útil) para resolver essa questão* (João is (important + useful) to solve this issue). Also, certain A_{pred} have subjunctive complement clauses introduced by *por* in European Portuguese which can be reduced to infinitive clauses when subordinate and matrix subjects are coreferential. These usually convey desires and their complement clauses always denote future propositions [Casteleiro 1981]. In BP, although *por* is acceptable, these seem to be more common with *para*: *Zé está (ansioso + louco + desesperado) (para viajar + para que Maria viaje)* (Zé is (anxious + crazy + desperate) (to travel + for Maria to travel)).

Ability A_{pred} are suggested to have prepositions in common in French (*à*) [Picabia 1978] and Italian (*di*) [Messina 2019]. In BP, these can be identified with a subset of A_{pred} accepting *em* (in) plus infinitive clause: *Ana é (hábil + eficiente + boa) em resolver problemas* (Ana is (skilled + efficient + good) in solving problems)

Certain adpositions have been noted to introduce “beneficiary” (or possibly “maleficiary”) complements of A_{pred} denoting human behavior [Nam 1996], [Carvalho 2007]. In BP, *com* (with) and its formal variety *para com* (roughly “towards”) seem to play this role: *Zé foi grosseiro (para + E) com Ana* (Zé was rude towards Ana).

4. Transformational features

A transformation in Lexicon-Grammar can be briefly defined as a regular relation between two paraphrastic sequences employing the same content morphemes [Harris 1968], [Harris 1991]. Common examples of transformations are passivization, relativization, and zeroing. We explore in this section transformations of A_{pred} pointed out as relevant for lexical semantics.

Appropriate nouns (N_{ap}) are those which can be erased in an argument without changing the information conveyed by the sentence. For example, *o comportamento da Ana é agressivo* (Ana's behavior is aggressive) and *a Ana é agressiva* (Ana is aggressive) mean roughly the same; in this case, *comportamento* (behavior) counts as an N_{ap} , since it is highly likely in this environment and may be zeroed with no loss of information [Harris 1991]. Acceptance of N_{ap} *comportamento* is distinctive of behavior predicates, but several other N_{ap} exist and classification based on them might lead to particularly fine-grained classes [Valetopoulos 2003]. Adjectives with N_{hum} subjects can be classified as relating to the different five senses based on their N_{ap} [Nam 1996]: *essa mesa é quadrada* (this table is square) = *o formato dessa mesa é quadrado* (the shape of this table is squared) conveys visual information; *esse queijo é fedido* (this cheese is stinky) = *o cheiro desse queijo é fedido* (the smell of this cheese is stinky) conveys olfactory information, etc.

Psychological A_{pred} generally correspond to sentence adverbs, whereas evaluative and behavior A_{pred} mostly correspond to either subject or object-oriented verb-modifying adverbs [Nam 1996]. This distinction can be verified by observing how well an adverb accepts a certain position. Psychological deadjectival adverbs can be sentence starters more easily: *infelizmente, o Zé respondeu à pergunta* (unfortunately, Zé answered the question); however, they sound like an interruption if closer to the verb: *?o Zé respondeu infelizmente à pergunta* (?Zé answered unfortunately the question). Evaluative and behavior deadjectival adverbs show the opposite behavior: *o Zé respondeu (corretamente + corajosamente) à pergunta* (Zé answered (correctly + bravely) the question) is perfectly acceptable, but *(?corretamente + corajosamente), o Zé respondeu à pergunta* ((correctly + bravely), Zé answered the question) have different overtones. This, however, does not seem to apply to A_{pred} with object clauses: e.g., *ansiosamente* (anxiously) is necessarily a verb modifier. Unacceptability of degree and manner adverbs was shown to be distinctive of nationality A_{pred} in European Portuguese [Carvalho 2007], which seems to also hold for BP.

Many intransitive A_{pred} can be characterized through their relation with nouns introduced by a possessive verb in relative clauses denoting a part-whole relation [Picabia 1978]: *barbado* (bearded) = *que tem barba* (that has a beard), *espinhoso* (thorny) = *que tem espinhos* (that has thorns). Both quality (e.g., *corajoso* “brave”) and material (*dourado* “golden”) A_{pred} have been suggested to correspond to certain constructions with V_{cop} plus preposition [Valetopoulos 2003]; in Portuguese, this sequence is *ser de* (to be of): *essa mulher é de uma enorme coragem* (this woman is of an enormous courage) = *essa mulher é muito corajosa* (this woman is very brave). Another type of nominalization distinguishes “disposition” A_{pred} , whose nominal version might occur in certain locative constructions [Valetopoulos 2003], such as with BP *invadir* (invade): *o cansaço invadiu Pedro* (tiredness invaded Pedro) = *Pedro ficou cansado* (Pedro got tired).

A correspondence between certain preposition plus clause segments and causative constructions has also been pointed out as a property of A_{pred} denoting psychological phenomena [Nam 1996] or “states of mind” [Messina 2019]: *Lia está contente de ir a Paris* (Lia is happy to go to Paris) = *Ir a Paris deixa a Lia contente* (Going to Paris makes Lia happy).

Aspectual pseudocopulas have been mentioned as widely accepted by disease A_{pred} , whereas only *tornar-se* (become) was characteristic of nationality and philosophical A_{pred} [Carvalho 2007]: e.g., *Zé (ficou + permanece + tornou-se) doente* (Zé (got + remains + became) sick), but *Zé (*ficou + *permanece + tornou-se) brasileiro* (Zé (*got + *remains + became) Brazilian). *Sentir-se* (feel) was observed to be equivalent to *estar* only in combination with “psycho-physiological” A_{pred} [Valetopoulos 2003], [Carvalho 2007]: *Zé se sente cansado* (Zé feels tired) = *Zé está cansado* (Zé is tired); however, *sentir-se* and *estar* lead to different meanings with non-psychological A_{pred} : *Zé sente-se prestes a conseguir um aumento* (Zé feels about to get a raise) ≠ *Zé está prestes a conseguir um aumento* (Zé is about to get a raise).

Certain verbs taking A_{pred} as arguments (find, consider) are said to apply only to psycho-evaluative adjectives [Nam 1996]. This is similar to BP *achar*, which is common with A_{pred} such as *suspeito* (suspect), *bom* (good), or even *alto* (tall), but seem less likely with *prestes a* (about to) or *sujeito a* (subject to).

“Object raising” or “tough movement” is the operation of moving the object of a complement clause to the subject position of the main clause, e.g. *É fácil agradar o Zé* (It is easy to please Zé) = *O Zé é fácil de agradar* (Zé is easy to please). Adjectives accepting this transformation are not associated to the same semantic properties by all of our sources: for French, they are said to be “appreciative” of the movable argument, whereas the ones that do not accept object raising are “descriptive” [Picabia 1978]; alternatively, these A_{pred} are said to denote the speaker’s “affective or emotional position” towards a proposition [Casteleiro 1981], in European Portuguese, or to evaluate both the action denoted by the subordinate clause and the raised object, in Korean [Nam 1996]. These A_{pred} have also been associated to the concept of difficulty in Italian [Messina 2019]. A large-scale study is required to pinpoint the semantic properties object raising might map to in BP.

A possibly overlapping transformational property is the acceptance of exclamative clauses such as *que espantoso!* (how astonishing!), *que horrível!* (how horrible!), etc., all of which convey the speaker’s emotions towards its (implied) subject. Additionally, it has been suggested that A_{pred} accepting this property are commonly factive, accept subjunctive clauses, and correspond to causative constructions: *espantoso* (astonishing) = *que causa espanto* (that causes astonishment) = *que espanta* (which astonishes). These are called “emotive adjectives” [Casteleiro 1981].

In European Portuguese over 50% of intransitive A_{pred} with N_{hum} subject may also be employed as nouns, most of which (80%) denote negative physical or psychological properties: *burro* (dumb), *idiota* (idiot), etc. Negative adjectives are also common after verbs like *chamar* (call, as in “Mary called him stupid”) and exclamative constructions with *seu* (literally “your”, interpreted as “you” in sentences such as “you bastard!”) [Carvalho 2007]. These constructions seem to have the same connotations in BP.

Adjectives denoting nationalities, beliefs, and diseases do not accept the pre-nominal position [Carvalho 2007]: **brasileira mulher* (Brazilian woman). The pre-nominal position is widely studied and usually associated to emotion and subjectivity in Brazilian grammars, but no comprehensive study such as those carried out in Lexicon-Grammar seems to have dealt with this subject in BP.

Intersective versus non-intersective adjectives is a well-known semantic division which has a formal counterpart: *um carro vermelho* (a red car) is intersective because it implies *algo que é um carro e é vermelho* (something which is a car and red); on the other hand, *um professor bom* (a good teacher) does not imply *alguém que é um professor e é bom* (someone who is a teacher and is good), but *alguém que é bom como professor* (someone who is good as a teacher). Shape (*oval*), nationality and social group (*grego* greek, *católico* catholic), and color (*vermelho* red) adjectives are said to be intersective, whereas denominal (*polar*), manner (*rápido* fast), emotion (*pobre menino* poor boy), and modal adjectives (*antigo* former) are non-intersective [Valetopoulos 2003].

Truth value A_{pred} can commonly be paraphrased with a different structure in which the adjective applies to the second element of a sentence with *ser* (be): *é possível que João seja professor* (it is possible that João is/will be a teacher) = *João é um possível professor* (João is a possible teacher) [Messina 2019].

5. Final remarks

This paper synthesized how observations from large-scale work on the syntax-semantics of predicative adjectives can provide promising features for a description of these predicates in Brazilian Portuguese. We have discussed 40 formal features: 8 distributional features applying to both arguments, totaling thus 16 features, plus two possibilities of copular verb, alternation of prepositions (counted here as a single property), number of arguments, and 20 transformational features. The exact role of all 40 properties in the semantics of Brazilian Portuguese is not completely clear, as the studies diverge on certain correlations and not all of them apply directly to Brazilian Portuguese. This study should thus be taken only as providing directions for a large-scale study of Brazilian Portuguese predicative adjectives, and not as an attempt of classification all by itself.

In addition, these are not the only properties that can be used to characterize predicative adjectives. Other recent references on Brazilian Portuguese have been discussing the semantics of formal properties applying to smaller sets of adjectives. Testing them in a larger lexicon might prove useful for adjective classification and their collection is recommended as a subsequent task. Also, irregularities are expected in how syntax and semantics relate, as it has been repeatedly reported in this type of study. Still, the features discussed in this paper presented substantial semantic correlations in other studies and can provide a solid foundation for a similar syntactic-semantic classification of Brazilian Portuguese predicative adjectives.

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.

References

- Barros, C. (2014) Descrição e classificação de predicados nominais com o verbo-suporte fazer no Português do Brasil. PhD thesis, Universidade Federal de São Carlos, São Carlos, Brazil.
- Calcia, N. (2022) Dar e receber um abraço: uma análise da conversão em português brasileiro. PhD thesis, Universidade Federal de São Carlos, São Carlos, Brazil.
- Cançado, M., Amaral, L. and Meirelles, L. (2018). Verboweb: Uma proposta de classificação verbal. In: *Revista da Anpoll*, n. 46, pages 123–141.
- Carvalho, R. (2007) Análise e representação de construções adjectivais para processamento automático de texto: adjectivos intransitivos humanos. PhD thesis, Universidade de Lisboa, Lisbon, Portugal.
- Casteleiro, J. (1981) Syntaxe transformational do adjectivo: regência das construções completivas. Instituto Nacional de Investigação Científica, Lisbon.
- Gross, G. (1994) Classes d’objets et description des verbes. In *Langages*, n. 115, pages 15–30.
- Gross, M. (1975). Méthodes en syntaxe. Hermann, Paris.
- Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. In *Langages*, n. 63, pages 7–52.
- Harris, Z. (1968) Mathematical structures of language. John Wiley & Sons, New York.
- Harris, Z. (1991) A theory of language and information: a mathematical approach. Clarendon Press, Oxford.
- Messina, S. (2019). The predicative adjective and its propositional arguments: a lexicon-grammar classification. In *Lingvisticae Investigationes*, n. 42 v. 2, pages 234–261.
- Nam, J. (1996) Classification syntaxique des constructions adjectivales en coréen. John Benjamins Publishing Company, Amsterdam.
- Picabia, L. (1978) Les constructions adjectivales en français : systématique transformationnelle. Librairie Droz, Geneva.
- Rassi, A. (2023) O verbo dar em português brasileiro: descrição, classificação e processamento automático. Letraria, Araraquara.
- Santos, M. (2015) Descrição dos predicados nominais com o verbo-suporte 'ter'. PhD thesis, Universidade Federal de São Carlos, São Carlos, Brazil.
- Valetopoulos, F. (2003) Les adjectifs prédictifs en grec et en français : de l'analyse syntaxique à l'élaboration des classes sémantiques. PhD thesis. Université Paris 13, Paris, France.

Vendler, Z. (1968) *Adjectives and Nominalizations*, Mouton, The Hague, Paris.

Building a Frame-Semantic Model of the Healthcare Domain: Towards the identification of gender-based violence in public health data

Lívia Dutra^{1,2}, Arthur Lorenzi¹, Lorena Larré¹, Frederico Belcavello¹, Ely Matos¹,
Amanda Pestana¹, Kenneth Brown¹, Mariana Gonçalves¹, Victor Herbst¹, Sofia
Reinach³, Renato Teixeira³, Pedro de Paula³, Alessandra Pellini⁴, Cibele Sequeira⁵,
Ester Sabino⁴, Fábio Leal⁴, Mônica Conde⁴, Regina Grespan⁵, Tiago Torrent^{1,6}

¹ Universidade Federal de Juiz de Fora (UFJF)

² Göteborgs Universitet (GU)

³ Vital Strategies Brasil

⁴ Universidade Municipal de São Caetano do Sul (USCS)

⁵ Secretaria da Saúde do Município de São Caetano do Sul

⁶ Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

{livia.dutra, arthur.lorenzi, lorena.tasca, mariana.goncalves,
victor.herbst, amanda.pestana, kenneth.cyrill}@estudante.ufjf.br,
{fred.belcavello, ely.matos}@ufjf.br, {sreinach, rteixeira,
pcbpaula}@vitalstrategies.org {acgpellini, cibelesqueira}@gmail.com,
sabinoec@usp.br, fabio.leal@inca.gov.br, monica.tilli@online.uscs.edu.br,
regina.maura@saocaetanodosul.sp.gov.br, tiago.torrent@ufjf.br

Abstract. Public data systems gather different information about Brazilian citizens. Such information is inserted in the system both via the selection of parameterized options and via open text fields. In this paper we describe the effort of modeling semantic frames for the lexicon of the healthcare domain as a means of tagging the open text fields in public health data to make them more easily interpretable by machine learning systems. This effort is one of the steps in a larger project aiming at using data science and machine learning techniques for the identification of territories prone to suffer from gender based-violence. The modeling effort currently covers 1,787 lexical units in the healthcare domain in Brazilian Portuguese, distributed in 29 semantic frames.

1. Introduction

According to the World Health Organization, one in three women¹ has been a victim of physical or sexual violence by their partner at some point in their lives. In Brazil, the notification of violence cases by healthcare professionals in SINAN² is mandatory. However, underreporting is a serious problem in tackling Gender-Based Violence (GBV). Reasons for underreporting described in the literature [GARBIN et al., 2015; KIND et al., 2013] include excessive workload of healthcare teams, lack of knowledge about the importance of the notification process, fear of possible retaliation by the aggressors, and, finally, difficulties identifying that the injuries and other health conditions are related to a violent episode. On the other hand, victims of GBV, when

¹ <https://www.who.int/publications/i/item/9789241564793>.

² SINAN is the Brazilian national information system for the notification of violence and diseases.

seeking healthcare services, may be included in other public health systems, such as e-medical records, SIM³ and SIH⁴.

Some of the Brazilian healthcare information systems feature, on top of parameterized data fields—that is, those where data is inserted via the selection of one option from a closed list—open text fields. So far, information present in those open text fields has been of little to no use in studies tackling GBV and other public health issues. The reasons for this lack of use of open text fields relate to the fact that language form can be ambiguous, polysemous and highly variable. Therefore, as a means to represent the semantics in the text and make the information present in open fields of public health systems more suitable for large scale data analyses, metadata can be associated to the text form. In this paper, we present the first part of an effort of modeling, in terms of Frame Semantics [FILLMORE, 1982] and using the FrameNet Brasil database structure [TORRENT et al., 2022], the lexical domains of Healthcare and Violence so as to represent the semantics of the linguistic forms present in the open text fields of national information systems.

The effort is part of a larger project aimed at using data integration and textual analysis to identify patterns that suggest that women registered in the health systems are victims of violence. Identification of patterns of GBV will be treated at the level of the territories where candidate victims live, and the resulting system will not keep present information on individuals. Hence, the main goal is to better equip policy makers, local authorities and health teams acting on said territories to design and apply public policies for both raising awareness and eventually reducing GBV in the territories. The hypothesis motivating the work presented in this paper is that data present in medical records, when linked to those in other databases and properly analyzed for their semantic content, can contribute to the identification of augmented risk of gender-based violence (GBV) at a given territory.

In this paper, we will report how frames in the Healthcare domain were modeled, from the first contact with the corpora, through their compilation, analysis of the lexicographic affordances of words and subsequent clusterization of them. We also approach the creation of frames from the most cohesive clusters, the establishment of relationships between frames and the association of lexical units to the frames created for lexicographic annotation.

2. The FrameNet Model

FrameNet⁵ is a lexicographic resource that originally applied the theory of Frame Semantics [FILLMORE, 1982] to the analysis of the lexicographic affordances of lexical items in English. From the original project, founded in 1997, other framenets have been developed for several languages, including Brazilian Portuguese [TORRENT et al., 2022].

The foundational principle behind any FrameNet analysis is the one according to which "meaning is relativized to scenes" [FILLMORE, 1977]. This is to say that, for every lexical item in any given language, the meaning of such an item is a function of a background scene defined in terms of the participants and props taking part in it. As an

³ SIM is the Brazilian mortality information system.

⁴ SIH is the Brazilian hospital admission system.

⁵ <https://framenet.icsi.berkeley.edu/>

example, consider a lexical item such as *arthritis.n*. To properly understand the meaning of this lexical item, we have to consider a scene—or frame—where an Ailment affects a Patient. Those two participants are necessary for the frame to be instantiated and, therefore, are the core frame elements (FEs) in the `Health_conditions` frame, shown in Figure 1. Other FEs may also be mentioned, such as the `Body_part` affected by the condition, or a `Symptom` of the Ailment. In Frame Semantics terms, the lexical unit (LU) *arthritis.n* evokes the `Health_conditions` frame.

Moreover, the background scenes evoked by LUs are connected to each other via a series of typed relations, forming a network of frames, or a FrameNet. For the `Health_conditions` frame, such relations model: (a) via Inheritance, that this frame is a more specific type of the `Gradable_attributes` and of the `State` frames; (b) also via Inheritance, that an `Epidemic` is a specific type of `Health_condition`; (c) via Using, that `Symptoms` and `Body_parts` may be required for understanding the conditions; (d) that the `Health_conditions` frame is referenced by frames modeling the notions of `Recovery` and `Cure`; and (e) that this frame is a part of `Healthcare_scenario`, among other relations.

In FrameNet methodology (RUPPENHOFER et al., 2016), frames are proposed based on a combination of domain knowledge and corpus analysis providing evidence of the valence affordances of LUs. For the project reported in this paper, a domain-specific corpus was used. We describe the corpus next.

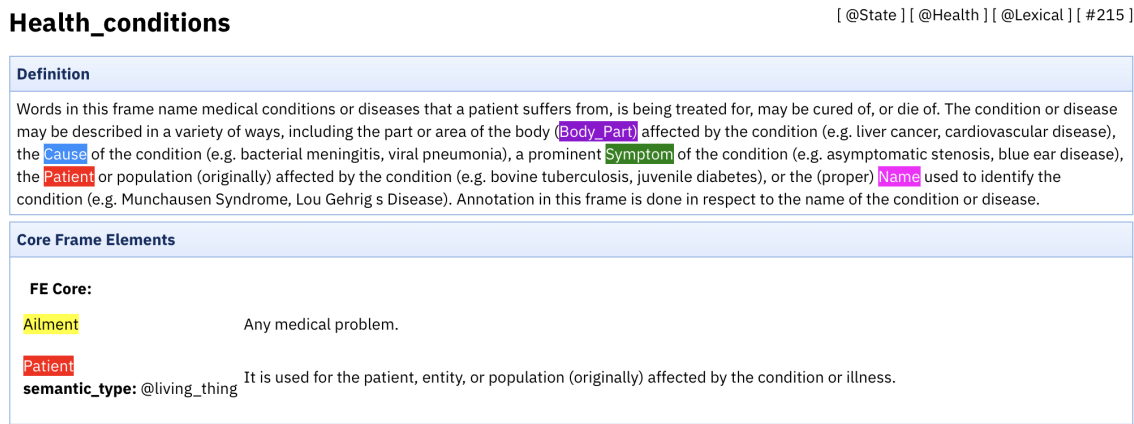


Figure 1. The `Health_conditions` frame

3. Materials

The methodology adopted for modeling the frames relies on the availability of a corpus of texts related to the Healthcare domain. The corpus used in this study was built using records extracted from the primary care information system—e-medical records—from the city of São Caetano do Sul in the state of São Paulo. For this study, the only piece of information that was used to build the corpus was the field in which health workers write down the patients' main complaints. These complaints are, in most cases, a single sentence stating the patient's symptoms and existing medical conditions, as well as for how long they have been occurring. Although considerably less common, some records contain other types of information, such as previous medical interventions, which in turn, can contain personally identifiable information (PII).

Taking into consideration the patients' rights to the privacy of their own data, especially when some could be victims of GBV, the most important steps in building the corpus were (a) extracting only the open text fields from the database, so that the lexicographers could not access the other data in the system, and (b) removing any PII from the texts. For this last process, any proper noun, dates and number were considered potential PIIs and excluded from the final corpus. All of the PII was replaced by tags in the text indicating that in the original text there was a name, date or number in that position. As a safeguarding measure, before any work was done with the corpus, one of the authors of this work⁶ manually checked that there were no remaining PII in the final corpus.

The final corpus contains 32,980 sentences and a total of 225,416 tokens, with an average of ~6.8 tokens per sentence.

4. Modeling the Healthcare Domain

The process of computational modeling for a specific domain begins with a general study of that domain, involving the identification of potential participants, events, and other essential elements. Costa (2020) presents a nine-step methodology for modeling a specific domain in FrameNet Brasil, as shown in Figure 2. This methodology was effectively used to model the Healthcare domain, apart from minor adjustments.

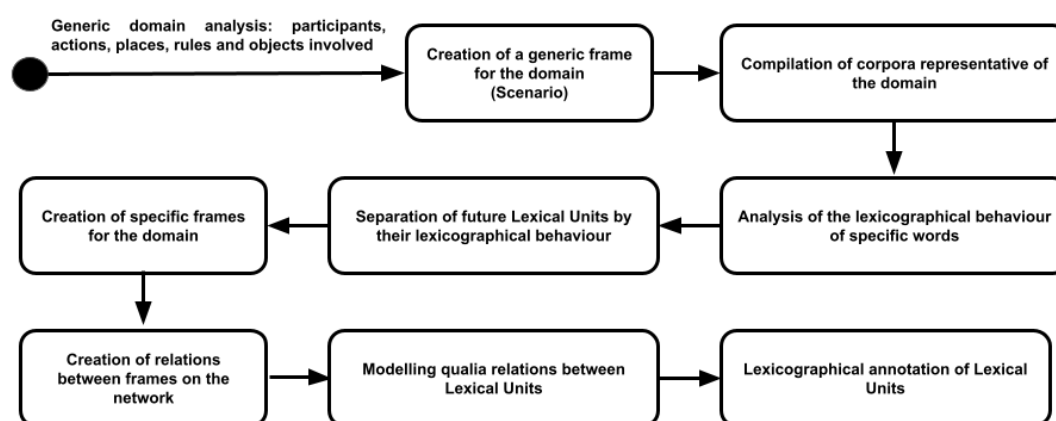


Figure 2. Steps in the process of computational modeling of a specific domain

The existing FrameNet database already included a generic frame called `Medical_interaction_scenario` along with 11 other frames intended for a potential Medical domain project. However, in order to capture a larger scope and cover the Healthcare domain, modifications to the FrameNet Brasil database were necessary. These adjustments were made in order to expand the frames beyond medical interactions and scenarios to include a more diverse range of scenes and participants. As a result, the generic frame was altered to `Healthcare_scenario`, and similar changes were made to the existing frames to align them with the Healthcare domain.

As predicted at the beginning of this work, these frames were not enough to cover the entire domain. Therefore, an analysis was made, grouping semantically related

⁶ Said author only had access to the corpus prior to the anonymization process in an encrypted computer environment with appropriate credentials, complying to the methodology registered and approved for the project by the National Research Ethics Committee (process number 64733922.3.0000.5083).

words from the corpus described in section 3, which identified the need to create new frames to meet the demands of the domain. In the following sections, we will discuss the details of these processes, as well as the other steps involved in modeling a domain.

4.1. Clusterization of Lexical Units

The first step of the methodology consists of finding a set of single and multi-word expressions as candidate LUs. These candidate terms are not proposed arbitrarily. Instead, they are extracted from the corpus described in Section 3. Using the 'Keywords & Terms' feature of the Sketch Engine software [KILGARRIFF et al, 2014a], the top scored 1000 single-word and the top 1000 multi-word expressions were selected. Scoring in the Keywords & Terms tool is based on the ratio between the expression frequency on the focus and on a reference corpus. The Portuguese Web 2020 Corpus was used as the reference corpus in this comparison.

The following phase was dedicated to the arrangement of the candidate LUs within an initial set of existing frames, which varied from frames that were not directly related to the Healthcare domain—e.g. the `Kinship` frame, which contains lexical units like *pai.n* 'dad', *mãe.n* 'mom', *irmão.n* 'brother'—and the existing frames associated to the Healthcare domain as the `Health_conditions` frame, shown in Figure 1.

From this point onwards, the remaining expressions were grouped and sorted according to their semantic similarity. As a result, domain-specific frames—such as the `Symptoms` frame, containing LUs like *náusea.n* 'nausea', *tosse.n* 'cough', and *choro excessivo.n* 'excessive crying'—were also created. The last example shows an interesting occurrence and showcases how this sorting process may vary due to the domain where the LU is inserted. Normally, LUs such as *choro.n* 'crying' and *excessivo.a* 'excessive' would both be considered and added into their respective frames individually: the first to the `Communication_Noise` and the second to the `Degree` frame, since the adjective describes the crying as excessive. However, within a specific domain, it is crucial to consider particularities such as considering the intensity marker as a part, and not as an accessory, of the LU which, thus, is added to the `Symptoms` frame.

As another example to illustrate the grouping and sorting process, take the `Health_intervention` frame, which consists of LUs that semantically represent intentional procedures performed to treat patients. Even if they differ a lot in magnitude—such as *curativo.n* 'bandage', *amputação.n* 'amputation', *exame.n* 'exam', *cirurgia.n* 'surgery'—, they are still within the same semantical boundary. As the number of candidate terms grew larger, so did the number of prototypical frames containing these terms, all sorted based upon their semantic function, which were then to be elaborated into frames *per se*.

4.2. Frame Creation Process

The frame creation process can follow the bottom-up, the top-down perspective, or a combination of both. The bottom-up approach consists in studying a corpus to create a frame, i.e., it takes into account linguistic evidence to structure the frame, analyzing patterns and relations. On the other hand, the top-down approach starts from the researcher's intuition, creating the frame structure as a starting point, almost a reverse process if compared to the first method. To model a specific domain both approaches

are combined [TORRENT et al., 2014]: we start with the bottom-up approach and check the intermediate analyses against the systematized knowledge of the domain.

In the bottom-up approach, the process begins with the previously mentioned grouping of lemmas derived from the selected corpus. Once this step is completed, the semantic and syntactic valence of the grouped lemmas is examined within the sentences in which they appear.⁷ This allows for the extraction of patterns, which helps to define and structure the frame. In parallel, the frame elements and their coreness are determined. Figure 3 shows the relevant information of a frame and demonstrates the structure of the created frame *Symptoms*.

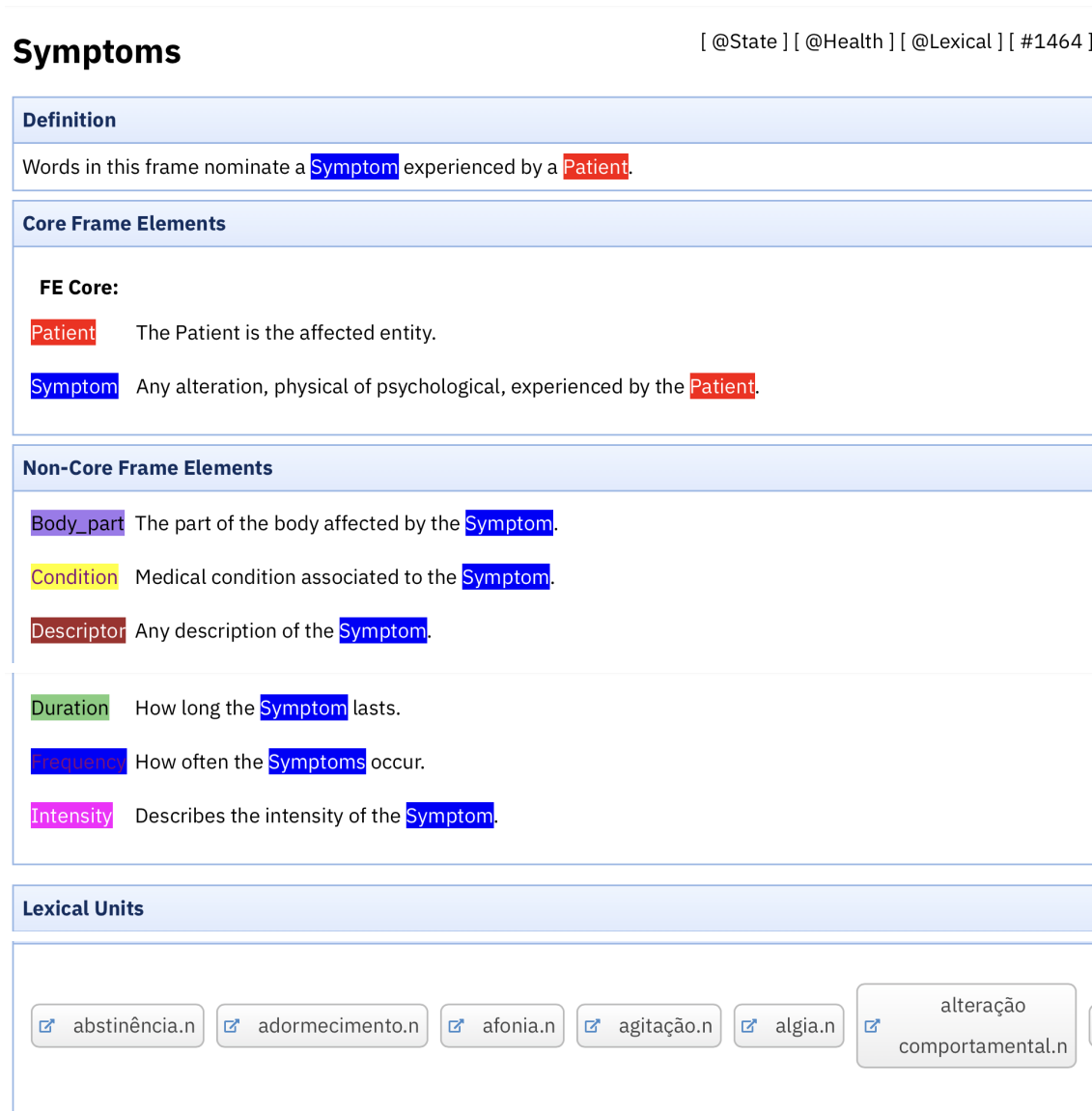


Figure 3. The *Symptoms* frame

Up to this point, 17 frames have been created to integrate the network that composes the Healthcare domain. Taking into account the existing frames that have been adjusted, the domain now consists of 29 frames with 1,787 lexical units associated

⁷ Access to the sentences in the corpora is limited to a selected number of researchers in the team, all of which are bound by non-disclosure policies established in the ethics protocol for the project.

with it. As already stated, the domain is a network—not a list—which means that the frames are interconnected at some level. This is established by means of frame relations.

4.3. Frame Relations

The existing frame-level relations in FrameNet Brasil can be of three types: between Frames (F-F), between Frame Elements (FE-FE) and between Frames and Frame Elements (F-FE) [TORRENT et al., 2022]. These relations not only allow for the organization of frames within a framework, but also improve the understanding of frames by providing additional semantic information.

The types of F-F relations are represented by arrows of different colors and were defined by Ruppenhofer et al. (2016) as: Inheritance (red), Subframe (blue), Precedes (black), Using (green), Causative_of (yellow), Incoative_of (brown), Perspective_of (pink) and See_also (purple).

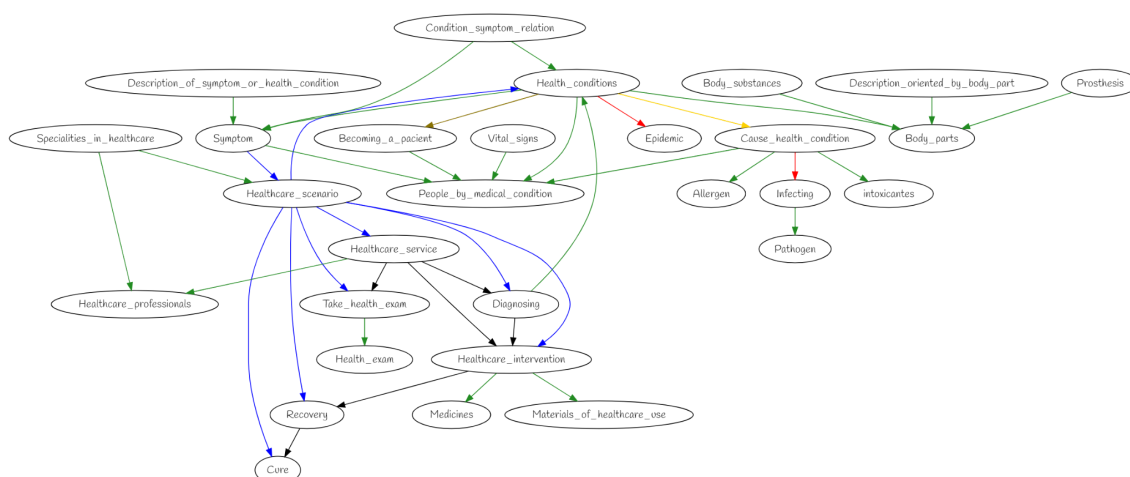


Figure 4. F-F Relations in the Healthcare Domain

Regarding FE-FE relations, it is possible to indicate similarities and correspondences between FEs from different frames. Within the Healthcare domain, for example, the Healthcare_Professional FE of the Healthcare_intervention frame is directly connected to the Professional FE of the Health_care_service frame. This connection indicates that the professionals engaged in the mentioned interventions are those related to healthcare. Furthermore, the F-FE relations, in a similar manner to the previous one, indicate a FE of a given frame that makes reference to another frame. This is the case, for instance, of the frame Take_a_health_exam which has the Health_exam FE that is directly related to the Health_exam frame. Such a relation points out that the exams covered by this FE are the LUs that fall under the Health_exam frame.

4.4. Qualia Relations

A different type of relation found in the FrameNet Brasil database structure are qualia relations based on the qualia structure proposed by Pustejovsky (1995). Unlike the relations discussed so far, qualia relations express associations between LUs. Qualia relations can be subdivided into four different types: agentive, constitutive, formal and

telic. This small set is, as expected, very general. To further specify these associations, FrameNet Brasil uses frames to mediate the relations between two LUs.

These types of connection are essential to capture aspects of meaning that cannot be adequately represented in the original FrameNet model. For instance, the LU *alergia.n* 'allergy' in the `Health_conditions` frame, can be associated to *amendoim.n* 'peanut' in the `Food_and_beverages` frame, as peanuts are known to cause allergic reactions in some individuals. This is an instance of the agentive quale that in FrameNet Brasil is also mediated by the `Cause_health_condition` frame, with *amendoim.n* 'peanut' filling the `Cause_of_the_health_condition` FE slot and *alergia.n* 'allergy' filling the `Health_condition` FE. This mediated association models the intuitive notion that when *alergia.n* 'allergy' evokes `Health_conditions`, frames such as `Food_and_beverages` may be relevant in the context.

4.5. Annotation

The *corpus* annotation represents the last step of the frame modeling process, and validates the model. For this project, following the methodology adopted by FrameNet Brasil, the full-text perspective was chosen for better results. In this approach, the researcher uses the entire *corpus* as a source, rather than predetermined words in selected sentences. Thus, all the meaningful LUs are annotated in relation to other constituents in the sentence. It allows for a much wider semantic and syntactic analysis upon the textual genre being annotated which aligns with the goals of the project.

Moreover, the annotated corpus, containing sentences in the Healthcare domain, the frames evoked in these sentences and the distribution of the relevant FEs, is an important resource to train semantically enriched machine learning models.

5. Conclusions

This paper has presented an overview of the effort to model the Healthcare domain according to FrameNet Brasil's methodology. It has shown a bottom-up methodology for semantic modeling, based on a corpus of the healthcare domain. The resulting network of frames, consisting of 29 frames, 17 of which are newly-created, already covers 1,787 LUs. This model of the domain was created as part of a larger project with the long-term goal of identifying traces and tendencies of GBV in certain territories in Brazil, based on public health system records. To achieve this goal, we plan on using, among other techniques, machine learning models. For future work, we propose the use of annotated data, connecting the texts written by health professionals to the frame structure created in this work, to enhance the quality of those models. A semantically enriched machine learning model is more likely to be informative (e.g. by also relating output to frames), but also has the potential of having better performance because it does not solely rely on raw texts. In this context, a better model is also essential to identify patterns that could be related to GBV, which could be in turn used by policy makers to take more informed actions.

Acknowledgements

Research presented in this paper was funded by the Data to Safeguard Human Rights Accelerator Program of the Patrick McGovern Junior Foundation.

References

- Costa, Alexandre Diniz. (2020) “A tradução por máquina enriquecida semanticamente com frames e papéis qualia.” (Ph.D. thesis in Linguistics. Universidade Federal de Juiz de Fora, Juiz de Fora.)
- Fillmore, Charles J.(1982) The case for case reopened. In: Grammatical relations. Brill, 1977. p. 59-81.
- Fillmore, C. J. (1982). Frame semantics. In: Linguistic Society of Korea (ed.), “Linguistics in The Morning Calm”. Seoul: Hanshin, p.111-138.
- Garbin, Cléa Adas Saliba et al. (2015) “Desafios do profissional de saúde na notificação da violência: obrigatoriedade, efetivação e encaminhamento.” In: Ciência & Saúde Coletiva, v. 20, p. 1879-1890.
- Kilgariff, Adam et al. (2014) “The Sketch Engine: ten years on.” In: Lexicography, v. 1, n. 1, p. 7-36.
- Kilgariff, Adam et al. (2014) “PtTenTen: A corpus for Portuguese lexicography.” In: Working with Portuguese Corpora, p. 111-30.
- Kind, Luciana et al. (2013) “Subnotificação e (in) visibilidade da violência contra mulheres na atenção primária à saúde.” In: Cadernos de Saúde Pública, v. 29, p. 1805-1815.
- Pustejovsky, James.(1998) The generative lexicon. MIT press.
- Ruppenhofer, Josef et al. (2016) Framenet II: Extended Theory And Practice. <https://Framenet2.Icsi.Berkeley.Edu/Docs/R1.7/Book.Pdf>).
- Torrent, Tiago Timponi et al. (2014) “Multilingual lexicographic annotation for domain-specific electronic dictionaries: The Copa 2014 FrameNet Brasil project.” In: Constructions and Frames, v. 6, n. 1, p. 73-91.
- Torrent, Tiago Timponi et al. (2022) “Representing context in framenet: A multidimensional, multimodal approach.” In: Frontiers in Psychology, v. 13. doi: 10.3389/fpsyg.2022.838441

Proceedings of 14th Symposium in Information and Human Language
Technology

TILIC 2023

VIII Workshop de Iniciação Científica em
Tecnologia da Informação e da Linguagem
Humana

Gramáticas Locais para Reconhecimento de Construções com Verbo Suporte em Português

Luís Enrique Santos Prado Vereau, Juliana Pinheiro Campos Pirovani

¹Departamento de Computação - Universidade Federal do Espírito Santo;
Alegre, ES - Brasil

luis.vereau@edu.ufes.br, juliana.campos@ufes.br

Abstract. *Natural Language Processing is an interdisciplinary subarea of Computer Science and Linguistics that aims to study the generation, representation and understanding of natural language by computers. This article describes the automatic generation of Local Grammars (LGs) for Supporting Verb Constructions (SVC) from six Lexicon-Grammar tables that describe a total of 468 CVS. For this, six parameterized graphs were created using the tool Unitex to extract information from these tables. The LGs generation was done through shell scripts and Unitex. The generated LGs were applied at the corpus aTribuna, where 211 structures in the form of the searched SVC were found.*

Resumo. *O Processamento de Linguagem Natural é uma subárea interdisciplinar da Ciência da Computação e Linguística que tem como objetivo o estudo da geração, representação e compreensão da linguagem natural por computadores. Este artigo descreve a geração semiautomática de Gramáticas Locais (GLs) para Construções com Verbo Suporte (CVS) a partir de seis tábuas do Léxico-Gramática que descrevem um total de 468 CVS. Para isto, foram criados seis grafos parametrizados utilizando a ferramenta Unitex para a extração de informações destas tábuas. A geração das GLs foi feita por meio de shell scripts e do Unitex. As GLs geradas foram aplicadas no corpus aTribuna, onde foram encontradas 211 estruturas no formato das CVS buscadas.*

1. Introdução

O Processamento de Linguagem Natural (PLN) é uma subárea interdisciplinar da Ciência da Computação e Linguística que se dedica a geração, representação e compreensão de linguagem natural de forma automática. Os próprios linguistas podem se beneficiar do PLN por meio das ferramentas construídas pelos profissionais da Computação, da mesma forma em que, simetricamente, a qualidade do PLN pode depender da descrição da língua pelos linguistas [Picoli et al. 2015]. [Chowdhury 2003] descreve PLN como sendo “uma área de pesquisa e aplicação que explora como computadores podem ser usados para entender e manipular texto e fala em linguagem natural para fazer coisas úteis”.

Devido a necessidade atual de compartilhamento e compreensão de informação na era da internet, essa área tem ganhado destaque cada vez maior no meio acadêmico e na indústria. A grande quantidade de informação disponível atualmente em textos de escrita livre (não estruturados) precisa ser tratada para uso em aplicações que buscam informações específicas a partir deles. Exemplos de aplicações relevantes que necessitam das técnicas de PLN são sistemas de perguntas e respostas, tradução automática e reconhecimento de entidades nomeadas.

Entre os diversos problemas encontrados pelas máquinas ao processar a linguagem natural, há a interpretação de expressões com significado não composicional, que não podem ter seu significado extraído analisando o sentido literal de suas palavras individualmente. Expressões Cristalizadas (EC) como “João comprou no mercado negro” e Construções com Verbo Suporte (CVS) como “João tem sangue frio para blefar” são exemplos dessas expressões. Este trabalho tem CVS como objeto de estudo, [Picoli 2020] expõe as diferenças entre EC e CVS.

CVS podem ser definidas como expressões compostas por um verbo que atua como Verbo Suporte (Vsup) e uma unidade predicativa não-verbal que pode ser um nome predicativo (Npred) como em “ter lábia”(Vsup+Npred), um adjetivo (adj) como na construção “estar liso”(Vsup+adj), ou uma expressão que se comporta como adjetivo (Expadj), “estar azul de fome”(Vsup+Expadj), por exemplo [Picoli 2020].

[Flores 2020] analisou ocorrências do verbo *dar* em duas coleções de textos em português brasileiro, o *corpus* do projeto Fala Goiânia¹ e o *corpus* do Grupo Discurso & Gramática². Notou-se que das 190 ocorrências do verbo no *corpus* do projeto Fala Goiânia, em 103 (54,21%) ele atuava como Vsup, e no *corpus* do Grupo Discurso & Gramática, das 70 ocorrências encontradas, em 47 (67,14%) o verbo assume papel de Vsup, indicando grande presença das CVS na língua portuguesa do Brasil.

Desta forma, o reconhecimento automático ou semiautomático de CVS é importante para o PLN pela riqueza de significado de seus elementos, a dificuldade de entendimento da não-composicionalidade pelo computador, bem como pelo frequente aparecimento de expressões desse tipo no português. A geração automática de resumos, tradução de máquina e *chatbots* [Tan et al. 2021] são possíveis tarefas nas quais o reconhecimento dessas expressões poderá ser aplicado.

Um dos métodos para descrever CVS é o Léxico-Gramática [Gross 1975] que consiste em criar tabelas, também chamadas de tábuas, que detalham um conjunto de expressões e possíveis variações para certos elementos, como verbo utilizado, ausência ou existência de negação, comparação, intensificação, dentre outras.

Uma forma de reconhecer expressões com características sintáticas e semânticas em comum, como as CVS, são as Gramáticas Locais (GLs). As Gramáticas Locais [Gross 1997] são “gramáticas de estados finitos ou autômatos de estados finitos que representam conjuntos de expressões de uma língua natural”.

Este trabalho tem como objetivo a extração de informações presentes nas seis tábuas do Léxico-Gramática que descrevem CVS em [Picoli 2020] para geração semi-automática de GLs que sejam capazes de reconhecer essas expressões e anotar seus significados. O trabalho busca, por fim, fornecer recursos que possam ser utilizados em aplicações de PLN e auxiliem tanto profissionais da computação quanto da linguística.

2. Metodologia

Unitex³ é um conjunto de software livres para PLN que permite, além do pré-processamento de textos, a construção de GLs e a construção automática de GLs a partir

¹<https://gef.letras.ufg.br/p/11947-projetos-tematicos>

²<https://discursoegramatica.wordpress.com/>

³<https://unitexgramlab.org/pt/>

de uma tabela (neste caso, tábua do Léxico-Gramática) e de um grafo parametrizado.

As GLs no Unitex são representadas por grafos como o apresentado na Figura 1 que reconhece a estrutura [Nome Próprio (reconhecido pelo código lexical <N+Pr >) + não (opcional) + verbo *ser*, *ter*, *permanecer* ou *continuar* (reconhecido pelo código lexical <Verbo.V >, como <ser.V >) + coração grande].

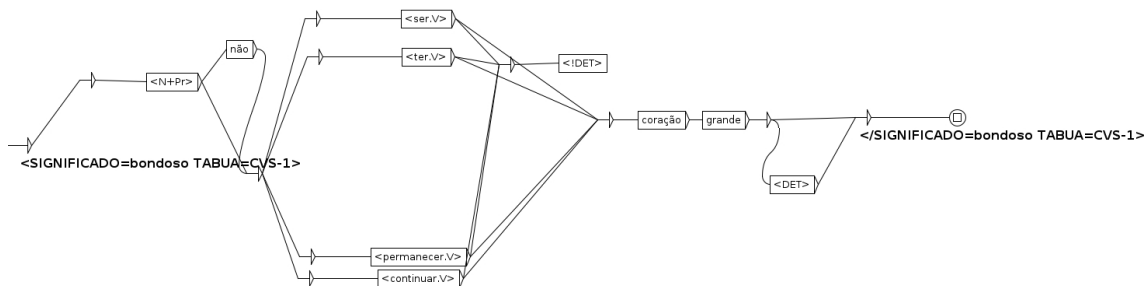


Figura 1. SubGL CVS-1

Um exemplo de tábua do Léxico-Gramática é apresentado na Figura 2. Na linha 3 desta tábua está descrita a CVS representada pela subGL da Figura 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989	990	991	992	993	994	995	996	997	998	999	1000
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	5																																																																																																																																																																																																																																																																																																																																																																																																																																											

As GLs geradas pelos grafos parametrizados foram aplicadas no *corpus* composto por um conjunto de textos publicados pelo jornal do Espírito Santo aTribuna⁴. Esse *corpus* contém textos jornalísticos de gêneros variados, tendo sido utilizado em [Santiago 2022]. Também foram criados *shell scripts* que automatizaram o processo de criação e aplicação das GLs ao chamar programas do Unitex responsáveis pela *tokenização* dos textos (Tokenize), aplicação de um Grafo Parametrizado em uma Tábua do Léxico-Gramática (Table2Grf), compilação das GLs (Grf2Fst2), dentre outros.

3. Resultados

Foram construídos 6 grafos parametrizados e geradas semiautomaticamente 473 GLs para reconhecimento de CVS. Para o *corpus* escolhido, foram encontradas 211 correspondências com estruturas no formato das CVS buscadas.

Algumas das expressões corretamente identificadas foram: "Vaz tem carta branca", "três é show de bola", "eu sou o máximo", "vida é um mar de rosas", "senão fica o dito pelo não dito", "público não esteve lá essas coisas", "Eduardo está entre a cruz e a espada", "Motta é do ramo" e "Sandra tem os dias contados", sendo "ter os dias contados" a CVS mais frequente, aparecendo 21 vezes (9,95%).

Observou-se também a presença de falso-positivos dentre as construções identificadas, i.e, expressões que foram identificadas como CVS, mas não são CVS. Por exemplo, anotou-se a expressão "o corpo foi achado", falso positivo para a CVS *ser um achado*, como em "o livro é um achado". Notou-se que a ocorrência deste falso-positivo em particular foi devido à descrição na tábua CVS-1cop que permite essas construções sem o artigo indefinido. Sabendo que este artigo é necessário para esta CVS, a tábua poderia ser alterada para corrigir essa situação melhorando a precisão da GL gerada.

Outra CVS que apresentou falso-positivos foi *ser o de menos*, que indica pouca relevância como em "essa informação será o de menos". Um dos falso-positivos identificados foi "A academia também observou que a taxa de acréscimo de novas reservas em todo o mundo era de menos", que indica que taxa de acréscimo de novas reservas em todo o mundo foi menor que certo valor, apresentando, desta forma, significado bem diferente ao da CVS.

Faz-se necessário, então, analisar os grafos parametrizados novamente a fim de mitigar a aparição de falso-positivos.

4. Conclusão

Neste trabalho foram geradas 473 GLs a partir de tábuas do Léxico-Gramática e grafos parametrizados construídos no Unitex. As GLs identificaram no *corpus* aTribuna 211 estruturas correspondentes a CVS.

Como metas futuras, os grafos parametrizados construídos até então sofrerão novas análises e serão feitas melhorias. Igualmente serão feitas novas análises nas tábuas do Léxico-Gramática a fim de remover possíveis falhas. Também será realizada a análise de resultados, avaliando-se a precisão das correspondências encontradas.

Pode-se avaliar também em quais gêneros textuais houve mais ocorrências de CVS e analisar se, nesses textos, há alguma CVS não descrita nas tábuas.

⁴<https://tribunaonline.com.br/>

Referências

- Chowdhury, G. (2003). Natural language processing. *Annual review of information science and technology*, pages 51–89.
- Flores, E. (2020). A construção-suporte no português brasileiro.
- Gross, M. (1975). Méthodes en syntaxe: régime des constructions complétives.
- Gross, M. (1997). The construction of local grammars. *Finite-state language processing*, pages 329–354.
- Picoli, L. (2020). Contínuo e limite entre expressão cristalizada e construção com verbo-suporte à luz do léxico-gramática.
- Picoli, L. et al. (2015). Uso de uma ferramenta de processamento de linguagem natural como auxílio à coleta de exemplos para o estudo de propriedades sintático-semânticas de verbos. *Linguamática*, pages 35–44.
- Santiago, D. (2022). Gramáticas locais para reconhecimento de expressões cristalizadas em português. *Relatório Parcial de Pesquisa, Programa Institucional de Iniciação Científica 2021/2022*.
- Tan, K. et al. (2021). Review on light verb constructions in computational linguistics. pages 25–26.

Anotação do Dataset Multimodal da ReINVenTA

Ana Carolina Loçasso Luz^{1,2}, Gabrielly Braz¹, Livia Pádua Ruiz^{1,2}, Mariane de Carvalho Pinto¹, Frederico Belcavello¹, Natália Sathler Sigiliano¹, Tiago Torrent^{1,2}

¹ Universidade Federal de Juiz de Fora (UFJF)

² Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

{livia.padua, ana.luz, gabrielly.braz,
mariane.carvalho}@estudante.ufjf.br, {fred.belcavello,
natalia.sigiliano, tiago.torrent}@ufjf.br

Abstract. *This paper aims to present an application of the semantic-computational model of FrameNet Brazil to the representation of multimodal objects. Therefore, it describes the steps involved in creating a subpart of the ReINVenTA Dataset, focusing on the semantic annotation of the TV series Pedro pelo Mundo for the modalities of text and dynamic images.*

Resumo. *Este artigo tem como objetivo apresentar uma aplicação do modelo semântico-computacional da FrameNet Brasil à representação de objetos multimodais. Para tanto, descreve as etapas envolvidas na criação de uma subparte do Dataset da ReINVenTA, com foco na anotação semântica da série de TV Pedro pelo Mundo para as modalidades de texto corrido e de imagens dinâmicas.*

1. Introdução

Ao longo da última década, pesquisas relacionadas aos campos da Visão Computacional e da Linguística têm feito crescer a popularidade de conjuntos de dados que combinam informações textuais e visuais – os chamados *datasets* multimodais [UPPAL et al., 2020]. Neste artigo, apresentamos os recentes esforços desenvolvidos com o objetivo de expandir os dados da FrameNet para o domínio multimodal [BELCAVELLO et al., 2020], além de especificar as aplicações do modelo semântico-computacional. Nesse sentido, é necessário compreender que as linguagens agem conjuntamente a fim de produzir sentido em um texto e, por isso, limitar a análise semântica dos objetos multimodais apenas a texto verbal seria prejudicial aos estudos [DÁNNELS et al., 2022]. Desse modo, são feitas anotações de texto corrido e de imagens dinâmicas referentes aos corpora constituídos.

2. A FrameNet Encontra a Multimodalidade

A FrameNet é um projeto lexicográfico computacional que tem como referencial teórico a semântica de *frames* proposta por Charles J. Fillmore (1982), a qual prevê um pareamento inerente entre empirismo e linguagem ao apontar que as palavras são representações de categorias de experiências e, portanto, evocariam “cenas” (*frames*) capazes de delimitar a maneira como interpretamos seu sentido. Assim, por *frame* entendemos uma representação esquemática de “qualquer sistema de conceitos relacionados de tal forma que, para entender qualquer um deles, é necessário compreender toda a estrutura que eles se encaixam” [PETRUCK, 1996]. Tal representação é formada através da experiência humana e capaz de relativizar o sentido de uma palavra a depender do *frame* evocado.

É ancorando-se na hipótese de Fillmore (1982) que a FrameNet investiga o ato da atribuição de sentido em línguas naturais. Até pouco tempo, o projeto havia concentrado sua atenção em apenas uma modalidade de análise semântica: a textual. Entretanto, ao considerarmos o aspecto inerentemente multimodal da comunicação humana [STEEN et al., 2018], vemos que contemplar as diferentes modalidades nas tarefas de anotação nos permite realizar uma análise semântica mais completa. Assim, a FrameNet Brasil (FN-Br), preocupando-se em promover essa análise e enriquecer o seu banco de dados, passou a buscar meios de abarcar essas diferentes modalidades, o que foi impulsionado no momento em que o projeto se integrou à Rede de Pesquisa e Inovação para Visão e Análise de Texto, a ReINVenTA. A pesquisa, que investiga o processamento semântico computacional de objetos multimodais, reúne diferentes laboratórios e grupos de pesquisa mineiros que trabalham na construção e avaliação de um modelo computacional para representar objetos multimodais [BELCAVELLO, 2023]. É nesse sentido que a FN-Br conta com uma proposta de anotação multimodal a partir de ferramentas próprias, que possibilitam um estudo acerca da interação entre diferentes modalidades da linguagem humana e de seu impacto na construção de sentido.

3. O Dataset Frame2

Partindo da concepção multimodal da comunicação humana e da hipótese de que os elementos visuais em um vídeo são capazes de evocar frames ou complementar o que foi evocado pela narração [BELCAVELLO et al., 2020], a FN-Br, por meio da formação de um dataset multimodal, produzido no âmbito da iniciativa ReINVenTA, busca fornecer uma forma de correlação entre os elementos visuais e textuais de uma produção audiovisual. Assim, tem-se como objetivo investigar a interação entre os frames anotados nas tarefas de anotação de texto corrido e nas de sequências de vídeo, de modo a comparar a maneira com que os frames mobilizados para os elementos textuais interagem em combinação entre áudio e vídeo.

Os objetos multimodais selecionados para a anotação de imagens dinâmicas foram os 40 episódios da série de viagens de TV "Pedro pelo Mundo", exibida a partir de 2016 no canal GNT. O programa é apresentado por Pedro Andrade e trata de aspectos sociais, culturais e econômicos dos diferentes países nos quais os episódios se passam. A primeira temporada, objeto de anotação que compõe o Frame2, conta com 10 episódios de 23 minutos cada. Cada episódio teve suas falas transcritas automaticamente e revisadas. Depois, os anotadores¹ ocuparam-se de anotá-las manualmente usando a *Web Annotation Tool* (WebTool)² – vide 4.1 –, conforme as diretrizes da FN-Br, ou seja, adotando uma abordagem perspectivizada para cada anotação. Em seguida, o mesmo anotador foi responsável por anotar, por meio da ferramenta de anotação multimodal Charon³ [BELCAVELLO et al., 2022] – vide 4.2 –, os elementos visuais presentes no

¹ O grupo de anotadores foi composto por graduandos em Letras, todos falantes nativos de português. Conforme explicitado por Belcavello (2023), 12 deles eram bolsistas do projeto, e outros 32 fizeram parte de oficinas de anotação oferecidas semestralmente pela equipe de pesquisadores da FN-Br na UFJF.

² Software de gerenciamento de banco de dados e anotação usado pela FrameNet.

³ A ferramenta foi desenvolvida para auxiliar na anotação de objetos visuais, na correlação desses objetos com dados textuais e na rotulagem dos frames e elementos de frame por eles evocados.

mesmo episódio, procurando guiar-se (mas não restringir-se) pelos frames e elementos de frame (EFs), identificados em cada trecho durante a anotação de texto corrido realizada anteriormente.

4. O Passo a Passo da Anotação

A FrameNet é um modelo semântico que tem seus itens lexicais organizados em Frames [FILLMORE, BAKER, 2009]. Por exemplo, o verbo *comer* evoca o Frame Ingestão, que, por sua vez, pressupõe a existência de um Ingestor e de um ou mais Ingeríveis. Além disso, Frames possuem uma cadeia de relações entre si. No caso, o Frame de Ingestão herda do Frame de Ingerir_substâncias e é usado por Alimentos_e_bebidas.

4.1. Anotação de texto corrido

Na tarefa de anotação de texto corrido, um grupo de anotadores recebe lotes de sentenças para análise, a qual consiste, primeiramente, em atribuir um frame a cada Unidade Lexical (UL) presente na sentença. Para isso, o anotador é orientado a clicar na UL que deseja anotar, de forma que, então, é carregado um quadro com todos os frames da FN-Br associados à UL correspondente. Após a escolha do frame, é gerada uma camada de anotação de EF, na qual é possível categorizar os demais itens da sentença. A anotação pode ser feita em mais camadas, mas, para fins deste artigo, apenas os EFs serão incluídos. Um exemplo pode ser encontrado na Figura 1, em que a UL *comer* foi anotada no Frame Ingestão e, a partir disso, pôde-se atribuir aos elementos *eu* e *um sanduíche de porco com molho caribenho*, respectivamente, os valores de Ingestor e de Ingeríveis.

Frame	Element	Value
[211691]	NI	Eu parei aqui para comer um sanduíche de porco com molho caribenho
Atividade_pausar.parar.v		Eu parei aqui para comer um sanduíche de porco com molho caribenho
FE	DNI DNI Ag	Luga Fina
Finalidade.para.prep		Eu parei aqui para comer um sanduíche de porco com molho caribenho
FE	DNI INC DNI Ag	Alvo
Ingestão.comer.v		Eu parei aqui para comer um sanduíche de porco com molho caribenho
FE	DNI DNI In	Luga
Alimentos_e_bebidas.sanduíche.n		Eu parei aqui para comer um sanduíche de porco com molho caribenho
FE	INC	Partes_constituintes
Alimentos_e_bebidas.molho.n		Eu parei aqui para comer um sanduíche de porco com molho caribenho
FE	INC	Parte

Figura 1. Anotação semântica para a Unidade Lexical *comer* no Frame Ingestão

4.2. Anotação de vídeo

A ferramenta de anotação de imagens dinâmicas conta com três painéis e com um arquivo de vídeo, detentor das entidades a serem anotadas, presente no canto superior esquerdo (Figura 2). Com isso, é concedida ao anotador a oportunidade de assistir ao contexto da sentença previamente anotada, o que pode ampliar, e até mesmo alterar, a sua perspectiva de anotação. Visto o vídeo, o anotador pode começar a atividade de marcação de objetos, que consiste na criação e edição de *bounding boxes* ao redor das entidades que se deseja anotar. É importante apontar que alguns objetos são criados

automaticamente pelo próprio *software* da ferramenta, mas é decisão do anotador mantê-las ou excluí-las. As atividades de criar objetos, rastrear-los, editá-los e excluí-los são feitas a partir dos respectivos botões localizados abaixo do arquivo de vídeo, junto aos botões de manipulação de reprodução do vídeo.

Após criar e editar a *bounding box* ao redor do objeto escolhido, o anotador atribui a ele um frame a partir de uma lista com todas as opções que constam na base de dados da FN-Br. Uma vez tomada essa decisão, é preciso atribuir um EF e um *Computer Vision Name* (CV Name). Essa categoria associa uma UL ao objeto delimitado, sendo essa qualquer UL do banco de dados da FN-Br que evoque um frame que estabeleça uma relação de herança com o frame de Entidade.

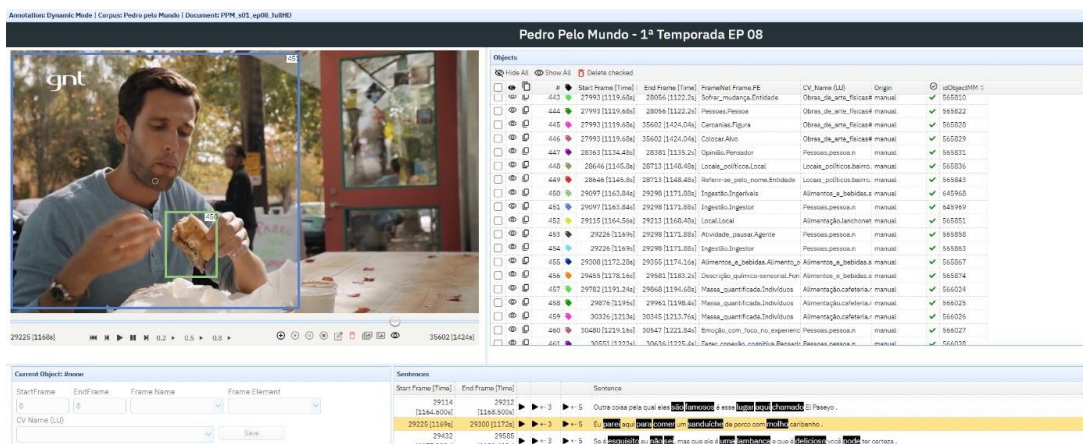


Figura 2. Interface de anotação de imagem dinâmica do corpus Pedro pelo Mundo

5. Resultados Alcançados

A partir de 2.195 sentenças, transcritas de 10 episódios do corpus Pedro Pelo Mundo, foi obtido o total de 11.796 *Annotation Sets* (AS) para texto corrido e 6.841 objetos para imagens dinâmicas. Uma vez que cada objeto visual é anotado para 3 categorias semânticas e que, em média, cada AS é anotado para 2,13 EFs, o esforço de anotação desse corpus produziu um *dataset* com 45.648 pontos de dados semânticos.

6. Considerações Finais

Neste artigo, descrevemos a proposta de anotação multimodal empregada pela FN-Br. Ao entender que a comunicação humana é um sistema multimodal, percebe-se a necessidade de se ir além de uma pesquisa que considera apenas a modalidade verbal para a construção de sentido. Assim, aponta-se a proposta de anotação da FN-Br como uma forma de realizar uma análise semântica de forma mais completa, visto que contempla, também, os aspectos visuais da comunicação.

Agradecimentos

A pesquisa apresentada neste artigo teve financiamento da FAPEMIG - processo RED-00106/21 e do CNPq - processos 408269/2021-9 e 420945/2022-9.

Referências

- Baker, C. F., Fillmore, C. J. and Lowe, J. B. (1998). "The Berkeley FrameNet Project". In: COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics.
- Belcavello, F.; Viridiano, M.; Diniz Da Costa, A.; Matos, E. E.; Torrent, T. T. (2020). "Frame-Based Annotation of Multimodal Corpora: Tracking (A)Synchronies in Meaning Construction". In: Proceedings of the LREC International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet. Marseille, France: ELRA, p. 23-30.
- Belcavello, F.; Viridiano, M.; Matos, E.; Torrent, T. T. (2022). "Charon: A FrameNet Annotation Tool for Multimodal Corpora". In: Proceedings of The 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022. Marseille, France: ELRA, p. 91-96.
- Belcavello, F. (2023). "FrameNet Annotation for Multimodal Corpora: devising a methodology for the semantic representation of text-image interactions in audiovisual productions". 135f. Tese (Doutorado em Linguística) — Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora.
- Dánnels, D.; Torrent, T. T.; Sigiliano, N. S.; Dobnik, S. (2022). "Beyond Strings of Characters: Resources meet NLP – Again". In: Volodina, E.; Dánnels, D.; Berdicevskis, A.; Forsberg, M.; Virk, S. (Org.). Live and Learn: Festschrift in honor of Lars Borin (pp. 29–36). Gothenburg: Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs Universitet.
- Fillmore, C. J. (1982). "Frame semantics". In: The linguistic society of Korea. Linguistics in the morning calm. Korea: Hanshin Publishing Company.
- Fillmore, C. J.; Baker, C. (2009). "A Frames Approach To Semantic Analysis". In: Heine, B.; Narrog, H. (Orgs.). The Oxford Handbook Of Linguistic Analysis (pp. 313–340). Oxford: Oxford University Press.
- Steen, F., Hougaard, A., Joo, J., Olza, I., Cánovas, C., Pleshakova, A., Ray, S., Uhrig, P., Valenzuela, J., Woźny, J. and Turner, M. (2018) "Toward an infrastructure for data-driven multimodal communication research". *Linguistics Vanguard*, Vol. 4 (Issue 1), pp. 20170041. <https://doi.org/10.1515/lingvan-2017-0041>
- Petruck, Miriam R. L. (1986) "Body Part Terminology in Hebrew: A Study in Lexical Semantics". Unpublished Ph.D. dissertation. University of California, Berkeley.
- Salomão, M. M. M. (2009) "FrameNet Brasil: um trabalho em progresso". *Calidoscópio*, [S. l.], v. 7, n. 3, pp. 171–182. Disponível em: <<https://revistas.unisinos.br/index.php/calidoscopio/article/view/4870>>. Acesso em: 6 ago. 2023.
- Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., & Zadeh, A. (2022). "Multimodal research in vision and language: A review of current and emerging trends". *Information Fusion*, 77, 149-171.

Pipeline para identificação de erros lexicais e geração de sugestões de correção

**Luana Q. Garcia¹, Miguel H. Chinellato¹,
Helena de M. Caseli¹, Leandro H. M. Oliveira²**

¹Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905 – São Carlos – SP – Brasil

²Empresa Brasileira de Pesquisa Agropecuária – Embrapa

{luanaqg,miguel.chinellato}@estudante.ufscar.br,

helenacaseli@ufscar.br, leandro.oliveira@embrapa.br

Resumo. No PLN, os textos são a principal fonte de informação na geração de modelos computacionais usando aprendizado de máquina. Entretanto, para que sejam úteis no processo de aprendizado, estes textos precisam representar corretamente o fenômeno que se deseja aprender e, neste caso, os erros lexicais podem ser impactantes. Este artigo apresenta a proposta de um pipeline para preparação e/ou correção de textos que identifica várias categorias de erros lexicais. O pipeline objetiva identificar, anotar e categorizar os erros contidos nos textos, bem como sugerir correções de forma automática.

1. Introdução

Na atualidade, a geração de modelos computacionais a partir de textos é realizada principalmente via Aprendizado de Máquina (AM). No *pipeline* de AM, a geração desses modelos passa por várias etapas das quais a de preparação de dados é a que demanda mais tempo [Chu et al. 2016], uma vez que modelos de alta performance exigem dados de qualidade para que se possa aprender padrões de alta qualidade [Zhang et al. 2003]. Embora a limpeza, correção e conversão dos dados sejam tarefas comuns e interdependentes [Ilyas e Rekatsinas 2022] em AM com dados textuais, o mesmo não se pode dizer sobre a descrição detalhada e ordenada dos passos realizados [Parulian e Ludäscher 2023].

Assim, tendo como referência algumas ferramentas de limpeza de dados para dados tabulares [Parulian e Ludäscher 2023, Li et al. 2023], a principal contribuição deste trabalho está na definição de um *pipeline* de preparação e/ou correção de textos (dados não estruturados) que pode ser aplicado a diversos domínios, operacionalizando esta tarefa e diminuindo débito técnico [Sculley et al. 2015] a longo prazo no modelo. Este artigo descreve a instanciação do *pipeline* para o domínio da agropecuária. O código do *pipeline* está disponível livremente¹.

2. O pipeline de correção textual

O *pipeline* de correção textual proposto neste trabalho identifica, anota e categoriza, de forma automática, os erros lexicais contidos nos textos. Para tanto, utiliza como recursos linguísticos: (1) um léxico do domínio geral composto por palavras do idioma sendo

¹Disponível em: <https://github.com/LALIC-UFSCar/pie-embrapa-pln>

processado e (2) uma lista de *stopwords* nesse mesmo idioma. A partir desses recursos, inicialmente, os *types* (palavras únicas) presentes nos textos são identificados e todos os *types* (exceto *stopwords*) não encontrados no léxico de domínio geral são classificados como palavras *desconhecidas*.

A partir do conjunto de *desconhecidas*, o *pipeline* identifica e categoriza as palavras em: (i) falta de acento, (ii) acentuação incorreta, (iii) erro de grafia e (iv) palavras aglutinadas². Sugestões de correção também são geradas. Outras categorias também são identificadas neste processo, tais quais: números, locais, termos específicos do domínio e palavras estrangeiras, mas estas não são consideradas erros.

Na identificação e categorização dos erros, o *pipeline* utiliza, além do léxico de domínio geral já mencionado, (3) uma base de dados de nomes de países, regiões, estados e municípios, (4) um léxico especializado do domínio, (5) um dicionário de frequências de palavras do idioma dos textos e (6) uma ferramenta de tradução para identificação de palavras em outras línguas. A Figura 1 ilustra os passos de categorização das palavras *desconhecidas* e seus respectivos recursos utilizados.

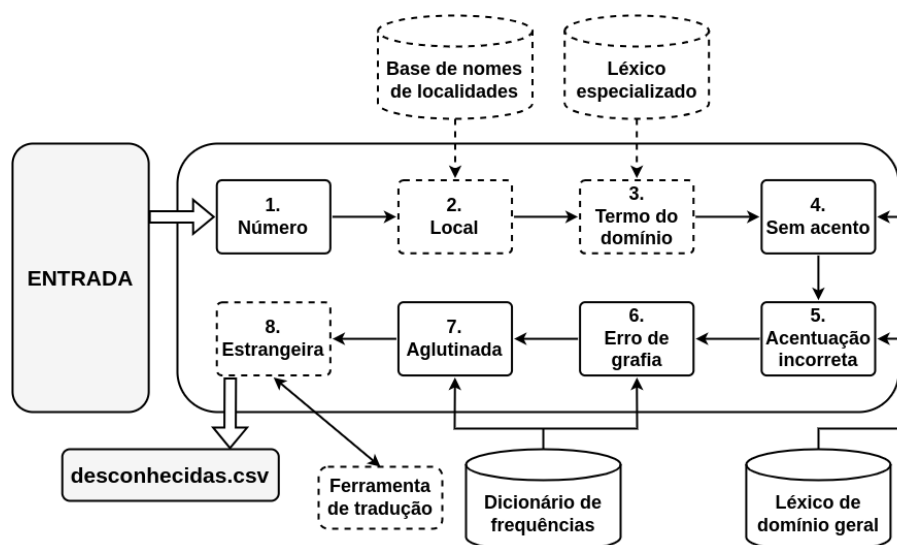


Figura 1. Pipeline de Categorização

A identificação e categorização dos erros ocorre na seguinte sequência de passos:

1. **verificação de números** – verifica se o *type* desconhecido é um número por meio do uso da função `isnumeric()` do Python, a linguagem de programação usada para implementação do *pipeline*;
2. **verificação de nomes de locais (opcional)** – identifica ocorrências de nomes de países, gentílicos, regiões, estados e municípios utilizando uma base de nomes de localidades;
3. **verificação de termos do domínio (opcional)** – verifica se a palavra desconhecida é um termo do domínio por meio da consulta a uma base de termos específicos da agropecuária, inseridos na língua portuguesa;

²Palavras que originalmente aparecem juntas no texto mas deveriam aparecer separadas.

4. **verificação de falta de acento** – verifica *types* que pertencem ao léxico de domínio geral da língua portuguesa, porém com todas suas palavras desacentuadas, identificando possíveis palavras sem acento;
5. **verificação de acentuação incorreta** – são removidos os acentos das próprias palavras *desconhecidas* para, em seguida, verificar se, agora, estas pertencem ao mesmo léxico desacentuado do passo anterior, identificando palavras com possíveis acentos errados;
6. **verificação de erros de grafia** – identifica erros de grafia por meio da biblioteca do algoritmo SymSpell³ para Python, que corrige palavras com pequenos erros de grafia através de distâncias de edição com base no dicionário de frequências de palavras do idioma desejado;
7. **verificação de palavras aglutinadas** – separa, com espaços em branco, palavras aglutinadas por meio de uma função⁴ que também utiliza o dicionário de frequência de palavras e assume que essas frequências seguem a Lei de Zipf;
8. **verificação de palavras estrangeiras (opcional)** – verifica se as palavras *desconhecidas* são possíveis palavras ou termos em outras línguas, nomeadamente o inglês, espanhol e francês, por meio de uma ferramenta de tradução.

A saída do *pipeline* é um arquivo de extensão csv intitulado *desconhecidas.csv* contendo 3 colunas: **palavra** (palavra identificada como desconhecida), **sugestão** (sugestão de correção do *type* desconhecido, nas categorias que são consideradas erros léxicos) e **categoria** (etapa do *pipeline* em que a palavra foi categorizada, sendo *desconhecida* caso nenhuma das categorias tenha sido detectada).

2.1. Definição da ordem de execução do *pipeline*

Diversos experimentos foram realizados em um corpus do domínio agropecuário composto pelos textos dos projetos da Embrapa para identificar a sequência dos passos previamente descritos, resultando na ordem ilustrada na Figura 1. Com o intuito de obter melhores resultados, optou-se por inserir nos passos iniciais do *pipeline* as verificações que se mostraram mais assertivas, notadamente: a verificação de números, de nomes de locais e dos termos específicos do domínio. Os próximos passos foram escolhidos de forma a, novamente, priorizar aqueles que demonstraram ser mais precisos durante testes empíricos. Desse modo, os passos seguintes são, respectivamente, as verificações de falta de acento e de acentuação incorreta. Ambos são bastante similares no quesito assertividade e funcionamento, utilizando o léxico desacentuado conforme descrito anteriormente.

Por fim, as últimas etapas do *pipeline* (antes da verificação de palavras estrangeiras) foram definidas como sendo, respectivamente, as verificações de erros de grafia e de palavras aglutinadas. Em ambos os casos, as funções retornam sugestões de correção corretas quando trata-se mesmo de um erro, com palavras como *principais* e *substituindo* sendo corrigidas, respectivamente, para *principais* e *substituindo* pelo SymSpell; e termos aglutinados como *comênfase* sendo corrigidos para *com ênfase* pela função de separação de palavras. Contudo, tais passos foram deixados mais perto do fim do *pipeline* pois apresentavam, com frequência,

³Disponível em: <https://pypi.org/project/symspellpy/>. Acesso em: 30 jun. 2023.

⁴Implementação baseada em: <https://stackoverflow.com/a/11642687>. Acesso em: 30 jun. 2023.

sugestões de correção para palavras que não eram erradas (como os termos específicos do domínio agropecuário). Por exemplo, o SymSpell sugeriu corrigir o termo apomíticos para apolíticos, assim como nematófagos para hematófagos, ambas palavras corretas porém particulares do domínio investigado. De maneira similar, a rotina de separação de *tokens* sugeriu corrigir negligenciamento para negligencia mento. Entretanto, muito mais numerosos eram os casos em que esta função de separação sugeria separar neologismos, como nanopartículas para nano partículas, agroativos para agro ativos e uma abundância de outros neologismos do contexto agropecuário. Escolheu-se deixar a verificação de palavras estrangeiras como o último do *pipeline* devido, sobretudo, a sua baixa assertividade⁵.

2.2. Instanciação do *pipeline* para o português do Brasil

Com intuito de exemplificar o uso do *pipeline* proposto, foi realizada a instanciação para processar textos escritos em português do Brasil no domínio da agropecuária contendo resumos de projetos da Embrapa. Para tanto, os recursos lexicais usados ao longo dos passos foram: **1. Léxico de domínio geral:** léxico POeTiSA (POrtuguese processing - Towards Syntactic Analysis and parsing)⁶; **2. Lista de stopwords:** *stopwords* da língua portuguesa do NLTK [Bird et al. 2009]⁷; **3. Base de nomes de localidades:** API do IBGE para Python⁸ que cobre os nomes de quase todos os municípios, estados e regiões do Brasil; e um arquivo JSON⁹ que contém os nomes e respectivos gentílicos de todos os países do Google Maps; **4. Léxico especializado:** AgroTermos¹⁰, que contém um conjunto de termos e conceitos específicos inter-relacionados semanticamente pertencente ao domínio agropecuário; **5. Dicionário de frequências:** dicionário de frequências de palavras da língua portuguesa extraído de todos os corpora disponibilizados pela Linguateca¹¹; **6. Ferramenta de tradução:** API do Google Tradutor para Python¹² objetivando a tradução de palavras da língua inglesa, espanhola e francesa.

3. Considerações finais

Tendo como foco a operacionalização do processamento de textos de um domínio específico, contendo possíveis termos especializados e erros de digitação/grafia produzidos por agentes humanos ou conversão automática de formato (por exemplo, de PDF para TXT), este artigo apresentou um *pipeline* de preparação e/ou correção de textos. Esse *pipeline* tem a grande vantagem de evidenciar os erros dos textos, separando o que é de fato erro e o que não é, além de oferecer sugestões para a correção humana.

⁵Nos experimentos realizados, notou-se que a API do Google Tradutor encontrava traduções errôneas para vários termos. Vale ainda ressaltar que esta é, com uma boa margem, o passo mais demorado de todo o *pipeline* uma vez que tal API funciona por meio de requisições Web, resultando em uma grande demora para retornar as traduções.

⁶Disponível em: <https://sites.google.com/icmc.usp.br/poetisa>. Acesso em: 30 jun. 2023.

⁷Disponível em: <https://www.nltk.org>. Acesso em: 30 jun. 2023.

⁸Disponível em: <https://pypi.org/project/ibge/>. Acesso em: 30 jun. 2023.

⁹Disponível em: <https://gist.github.com/jonasruth/61bde1fcf0893bd35eea>. Acesso em: 30 jun. 2023.

¹⁰Disponível em: <https://sistemas.sede.embrapa.br/agrotermos/>. Acesso em: 30 jun. 2023.

¹¹Disponível em: https://linguateca.pt/acesso/info_freq_English.php. Acesso em: 30 jun. 2023.

¹²Disponível em: <https://pypi.org/project/googletrans/>. Acesso em: 30 jun. 2023.

Agradecimentos

Este trabalho é resultado do Projeto Integrador Extensionista (PIE) vinculado ao Projeto de Extensão da UFSCar “Geração de representações de palavras no domínio agropecuário para melhoria dos processos de gestão da informação e conhecimento da Embrapa” (23112.035676/2022-76). Esse projeto é uma parceria com a Embrapa no projeto “Infopasto: Gestão da Informação e do conhecimento como suporte à gestão estratégica do Portfólio de Pastagens da Embrapa” (40.18.03.058.00.00) financiado pelo CNPq, a quem agradecemos o suporte financeiro.

Referências

- Bird, S., Klein, E., e Loper, E. (2009). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O’Reilly, 1st edition.
- Chu, X., Ilyas, I. F., Krishnan, S., e Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD ’16*, page 2201–2206, New York, NY, USA. Association for Computing Machinery.
- Ilyas, I. F. e Rekatsinas, T. (2022). Machine learning and data cleaning: Which serves the other? *J. Data and Information Quality*, 14(3).
- Li, P., Chen, Z., Chu, X., e Rong, K. (2023). Diffprep: Differentiable data preprocessing pipeline search for learning over tabular data. *Proc. ACM Manag. Data*, 1(2).
- Parulian, N. N. e Ludäscher, B. (2023). Trust the process: Analyzing prospective provenance for data cleaning. In *Companion Proceedings of the ACM Web Conference 2023, WWW ’23 Companion*, page 1513–1523, New York, NY, USA. Association for Computing Machinery.
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., e Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’15*, page 2503–2511, Cambridge, MA, USA. MIT Press.
- Zhang, S., Zhang, C., e Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17:375–381.

Coleta, composição e etapas de pré-processamento de *corpus*: procedimentos para a anotação multimodal da FrameNet Brasil

Anna Beatriz C. Silva¹, Iasmin Rabelo¹, Igor M. Oliveira¹, Mariana Souza¹, Maucha Gamonal², Raquel Roza¹

¹Laboratório de Tradução (LETRA) — Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG - Brasil

²Laboratório de Linguística Computacional (FrameNet Brasil)
Universidade Federal de Juiz de Fora (UFJF) — Juiz de Fora, MG — Brasil

bbangz@ufmg.br, {igormoliveira7, iasminvaleria,
marianassouza.mota, mauchaandrade,
figueiredorozar}@gmail.com

Resumo: Este trabalho apresenta a preparação de um *corpus* voltado para a anotação multimodal na FrameNet Brasil. A anotação, desenvolvida a partir da teoria da Semântica de *Frames*, permite a integração de diferentes modos comunicativos, construindo uma base de tecnologia linguística aplicável a múltiplas áreas. As etapas de coleta, composição e pré-processamento do *corpus* são os primeiros passos para o desenvolvimento das pesquisas de anotação multimodal.

1. Introdução

A integração da multimodalidade dentro do panorama teórico-metodológico da rede semântico-computacional da FrameNet [RUPPENHOFER *ET AL* 2016] amplia as possibilidades de desenvolvimento de tecnologias linguísticas mais avançadas através da base teórica da Semântica de *Frames* [FILLMORE 1982]. Isso envolve não apenas a análise e o processamento de texto, mas também a capacidade de interpretar e utilizar informações provenientes de outras modalidades, áudios, vídeos e imagens que são transcritos, contribuindo para aplicações mais sofisticadas em áreas, como, por exemplo, a tradução automática, a análise de sentimentos e a indexação de mídia.

Dentro desse cenário, o projeto em andamento ReINVenTA (*Research and Innovation Network for Visual and Textual Analysis of Multimodal Objects*) tem avançado nos processos de anotação de imagens estáticas e vídeos por meio das ferramentas Webtool e Charon [BELCAVELLO *ET AL* 2022]. A iniciativa reforça a representação de contexto a partir de uma abordagem multidimensional e multimodal [TORRENT *ET AL* 2022]. Liderado pelo laboratório de Linguística Computacional FrameNet Brasil na Universidade Federal de Juiz de Fora, o projeto conta com a colaboração de outras instituições, como a Universidade Federal de Minas Gerais.

Partindo disso, este trabalho apresenta as etapas da anotação estrutural [ALUÍSIO e ALMEIDA 2021] de um *corpus* multimodal, que alinha as transcrições textuais aos trechos de áudio/vídeo através de marcadores temporais [XIAO 2010]. Por

meio do *corpus Audition*, esta pesquisa visa descrever os estágios envolvidos na coleta, composição e etapas de pré-processamento de um *corpus* multimodal para anotação segundo o suporte teórico-metodológico da Semântica de *Frames* e da FrameNet Brasil.¹ Esses procedimentos antecedem a anotação multimodal com imagem dinâmica. Para tanto, apresentaremos a plataforma Charon, utilizada na realização de tais tarefas de pesquisa.

A seguir, na seção 2, apresentaremos a Semântica de *Frames* e como ela é usada pela FrameNet Brasil para realizar as análises multimodais. Na seção 3, exploraremos a metodologia utilizada na pesquisa e concluiremos o trabalho na seção 4, ao fazer ponderações sobre o trabalho realizado e as futuras perspectivas para o projeto.

2. Semântica de *Frames* e FrameNet Brasil em uma abordagem multimodal

A Semântica de *Frames* é uma teoria linguística desenvolvida por Charles J. Fillmore que propõe o estudo do significado a partir de uma perspectiva empírica e cognitiva. A teoria tem como base o conceito de *frame*, definido pelo autor como:

“[...] qualquer sistema de conceitos relacionados de tal modo que, para entender qualquer um deles, é preciso entender toda a estrutura na qual se enquadram; quando um dos elementos dessa estrutura é introduzido em um texto, ou em uma conversa, todos os outros elementos serão disponibilizados automaticamente.” [FILLMORE 1982, p. 111]

Dessa forma, a Semântica de *Frames* analisa como o conhecimento é “evocado” e expresso através da linguagem. Sob essa ótica, a compreensão de uma unidade lexical (UL) depende, necessariamente, da compreensão da cena, ou *frame*, em que ela está inserida. Visto que a teoria busca explorar a construção do significado considerando fatores sociais, culturais e cognitivos, surge a possibilidade de usá-la como auxílio no desenvolvimento de tecnologias linguísticas.

A FrameNet Brasil usa sua rede semântico-computacional para representar o conhecimento semântico e processar a linguagem. A multimodalidade, que abrange imagens, gestos e expressões faciais, enriquece o entendimento do significado, o que amplia o escopo do trabalho semântico-computacional. A abordagem multimodal da FrameNet Brasil, ao incorporar tipos de dados, como vídeos, programas de TV, imagens e textos [TORRENT ET AL 2022], oferece uma compreensão precisa do significado, considerando a interação complexa entre a linguagem e outras formas comunicativas.

3. Metodologia

Duas plataformas são utilizadas na anotação multimodal: a Webtool e a Charon. Na primeira, acontece a anotação linguística e, na segunda, a anotação de imagens estáticas e dinâmicas. O passo a passo para garantir que a anotação aconteça requer uma série de procedimentos, primeiramente, via Charon. Essa plataforma online é acessada mediante inscrição prévia. Tendo feito o login com usuário e senha, serão exibidas duas abas, “*Corpus*” e “*Annotation*”. Em “*Corpus*”, os *corpora* que compõem a pesquisa podem

¹ Agradecemos, pelo financiamento através de bolsas de iniciação científica, à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig) (PIBIC RED-00106/21) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (PIBIC 408269/2021-9 e 420945/2022-9)

ser acessados. Por meio desse ícone, estão os recursos que garantem as tarefas prévias à anotação em si. Nessa etapa, são inseridas informações básicas de identificação do material de análise, bem como o *upload* e pré-processamento do vídeo (Figura 1).

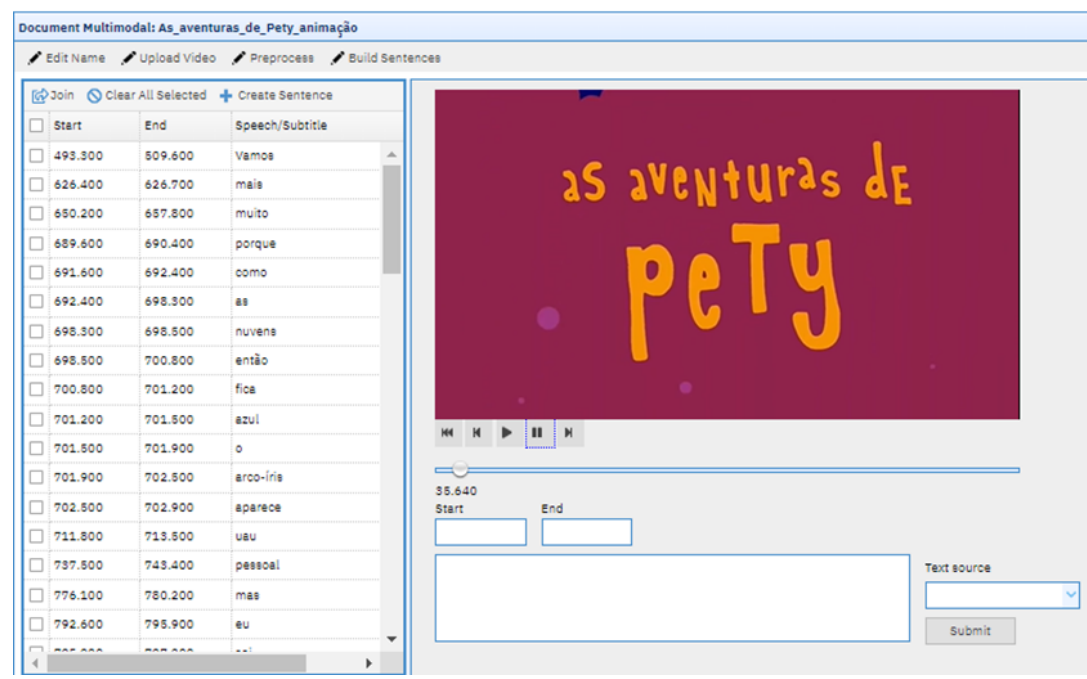


Figura 1. Página para a Edição das Sentenças do Curta-metragem

O *corpus Audition* inclui curtas animados e *live actions* com audiodescrição, como "As Aventuras de Pety", produzida por Aranhas Films, adicionado com a permissão da produtora e visível nas Figuras 1 e 2. A parte de texto corrido do *corpus* é composta pela extração dos áudio dos vídeos, o áudio é processado por um serviço em nuvem, que transcreve o que é dito e indica os intervalos de tempo. As legendas dos vídeos são capturadas e sincronizadas usando reconhecimento óptico de caracteres. O texto resultante do processo é exibido em um painel ao lado esquerdo da tela.

A Figura 1 mostra a etapa de revisão das sentenças reconhecidas automaticamente pela opção "**Build Sentences**". Nessa etapa, os anotadores acessam as informações e revisam o texto manualmente, fazendo correções e ajustes nos intervalos de tempo. Se necessário, eles podem criar novas sentenças ("**Create Sentence**"). As palavras são selecionadas no painel esquerdo e adicionadas ao painel direito através do comando "**Join**". Correções adicionais podem ser feitas na caixa de texto no centro da tela, como ilustra a Figura 2.

O tempo de início e fim das sentenças são ajustados usando as caixas "**Start**" e "**End**" acima da caixa de texto. O recurso "**Text source**" permite que o anotador classifique o texto que está sendo selecionado na caixa em "**Original audio**", para falas originais do vídeo, "**Subtitle**", para legendas, "**Text overlay**", para textos que aparecem na tela e não fazem parte da legenda, e "**Audio description**", para audiodescrição. Se um vídeo contém, por exemplo, áudio original, audiodescrição e legendas, três anotações serão feitas, cada uma com sua devida etiqueta. Para salvar as alterações, os anotadores selecionam a opção "**Submit**".

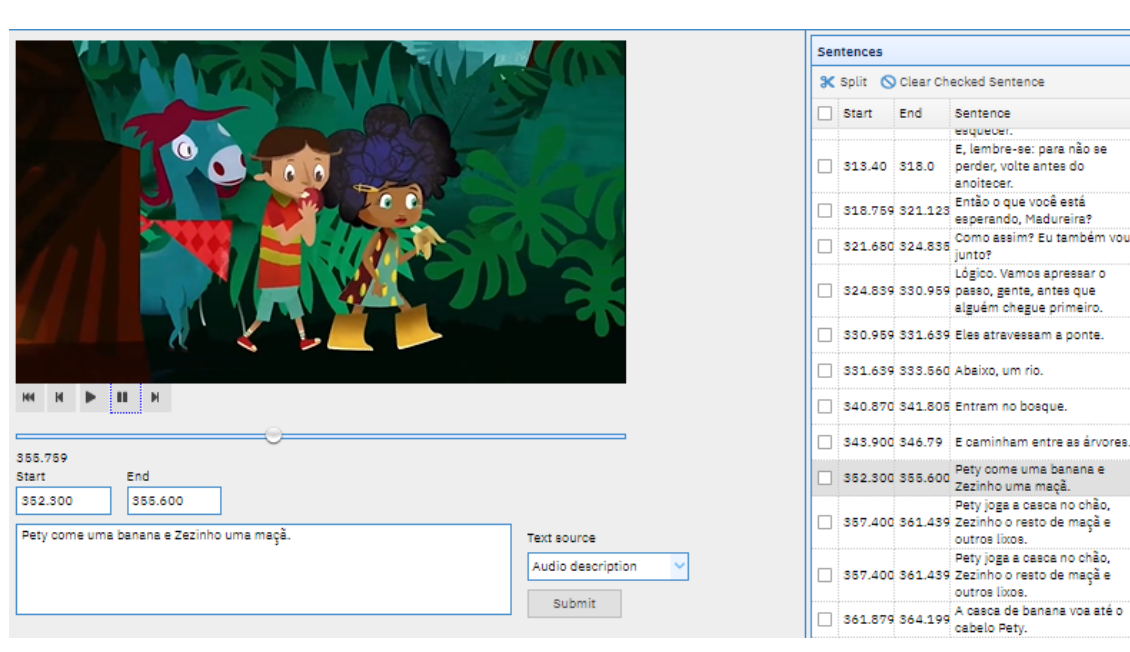


Figura 2. Espaço para Edição das Sentenças do Curta-metragem

Após o processo de transcrição do curta-metragem, passamos para a anotação de texto corrido, que acontece pela plataforma Webtool, onde fazemos as análises das sentenças e atribuímos os *frames* e elementos de *frame* ao texto. Feita a anotação linguística, fazemos a anotação do vídeo, anotação de imagem dinâmica, associando os *frames* evocados no texto, aos *frames* dispostos sobre o vídeo. As imagens são delimitadas por quadrados que as acompanham *frame* por *frame*, e são anotadas conforme o *frame* evocado, cv name (UL que identifica o objeto no vídeo) e unidade lexical.

4. Conclusão

A incorporação da multimodalidade no contexto da FrameNet Brasil proporciona um avanço significativo nas tecnologias linguísticas. Ao integrar informações de diferentes modalidades, – áudio, imagem e texto – é possível obter uma compreensão mais abrangente e precisa do significado na linguagem natural. Isso abre novas possibilidades de aplicações avançadas em áreas como tradução automática, análise de sentimentos e treinamento de IAs (Inteligências Artificiais). A anotação multimodal de *corpora*, por meio das plataformas Webtool e Charon, desempenha um papel fundamental nesse processo, permitindo a anotação de informações linguísticas, imagens estáticas e dinâmicas.

Com base nos avanços obtidos e no trabalho realizado com o *corpus Audition*, são planejadas ações para explorar a variedade de materiais analisados e avançar no desenvolvimento de tecnologias linguísticas. A área da audiodescrição, por exemplo, é grande beneficiária das pesquisas de anotação multimodal, visto que a associação entre o conteúdo linguístico verbal e não verbal atuam diretamente na experiência do usuário.

5. Referências

- Aluísio, S. M. e Almeida, G. M. de B. (2021). “O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística”, *Calidoscópio*, 4(3), p. 156–178. Disponível em: <https://revistas.unisinos.br/index.php/calidoscopio/article/view/6002>. Acesso em: 1 de jul. 2023.
- Belcavello, F., Viridiano, M., Matos, E. E. d. S., e Torrent, T. T. (2022). Charon: a FrameNet Annotation Tool for Multimodal *Corpora*. In *Proceedings of the 16th Linguistic Annotation Workshop*, páginas 91–96, Marseille, France, June. European Language Resources Association (ELRA). Disponível em: <http://www.lrec-conf.org/proceedings/lrec2022/workshops/LAWXVI/pdf/2022.lawvi-1.11.pdf>. Acesso em: 28 de jun. 2023.
- Charon [FNBr]. Disponível em: <https://charon.frame.net.br/>. Acesso em: 21 de jun. 2023.
- Fillmore, C. J. Semântica de Frames. In *Cadernos de Tradução*. Porto Alegre, nº 25, jul-dez, 2009.
- Pety; As Aventuras de. Direção: Anahí Borges. Produção: Anahí Borges. YouTube. 16 de jun. 2021. 14 min. Disponível em: <https://www.youtube.com/watch?v=h0TbaPIDkFI>. Acesso em: 21 de jun. 2023.
- Torrent T. T., Matos E.E. dos S., Belcavello F., Viridiano M., Gamonal M.A., Costa A.D. da, e Marim M.C. (2022). Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Front. Psychol.*
- Webtool [FNBr]. Disponível em: <https://webtool.framenetbr.ufjf.br/>. Acesso em: 21 de jun. 2023.
- Xiao, Richard. Empirical and Statistical Approaches. In: *Handbook of Natural Language Processing*. Nova York, 2ª Edição, 2010, p. 161.

Abordagens Baseadas em Léxicos para a Classificação de Sentimentos Orientada aos Alvos de Opinião em Comentários do Domínio Político

Lucas Lazarini¹, Fábio S. Igarashi Anno¹, Eloize R. Marques Seno¹, Helena M. Caseli²

¹Instituto Federal de São Paulo (IFSP) – São Carlos, SP

²Departamento de Ciência da Computação – Universidade Federal de São Carlos (UFSCar)
São Carlos, SP

{lazarini.lucas, fabio.seyiji}@aluno.ifsp.edu.br,

eloize@ifsp.edu.br, helenacaseli@ufscar.br

Resumo. *O enorme volume de textos opinativos produzido nas mídias sociais têm levado a uma busca cada vez maior por algoritmos capazes de analisar os sentimentos de pessoas em relação à produtos, entidades políticas, etc. Muitos modelos de análise de sentimento (AS) foram propostos para o português nos últimos anos. Contudo, a maioria deles consiste em analisar o sentimento geral de uma sentença, não considerando, portanto, o sentimento individual relacionado a cada alvo de opinião no texto. Dado este contexto, este artigo investigou o uso de léxicos de sentimentos na classificação de sentimento orientado ao alvo de opinião em comentários sobre debate político em português.*

1. Introdução

O crescimento explosivo das mídias sociais tem possibilitado aos seus usuários expressarem com facilidade opiniões e emoções/sentimentos sobre produtos, eventos políticos, indivíduos etc. por meio de *blogs*, fóruns de discussão e redes sociais. Como consequência, algoritmos capazes de analisar o sentimento e a opinião pública compartilhada por meio de textos opinativos têm ganhado cada vez mais importância. A Análise de Sentimentos (AS) é uma subárea do Processamento de Língua Natural (PLN), que utiliza ferramentas e recursos linguísticos-computacionais para analisar sentimentos em textos a respeito de indivíduos, produtos, serviços, entre outros [Liu and Zhang 2012].

Nos últimos anos muitos trabalhos foram proposto na literatura sobre AS. Contudo, no que se refere à língua portuguesa, a maior parte deles consiste em analisar o sentimento em relação à polaridade (positiva, negativa ou neutra) geral conduzida por um documento ou sentença (vide, por exemplo, [Chaves et al. 2012, França and Oliveira 2014, Capellaro and Caseli 2021]), não considerando as opiniões particulares relacionadas a cada entidade presente no texto, chamadas de alvos de opinião. Por exemplo, no comentário “Foi de facto um debate cordato, civilizado, em que Jerónimo se mostrou um senhor e o Louçã meteu a viola no saco.”, extraído do corpus utilizado nesta pesquisa (vide Seção 3), o sentimento em relação ao alvo “Jerónimo” é positivo, enquanto o sentimento em relação ao alvo “Louçã” é negativo.

Dado o contexto apresentado, este trabalho tem por objetivo investigar o uso de abordagens baseadas em léxicos na classificação de polaridade orientada para os alvos

de opinião em comentários sobre debate político em português. A escolha do domínio político é motivada pela quantidade limitada de trabalhos nesse domínio para o português.

O restante deste artigo está organizado da seguinte forma. A seção 2 descreve os principais trabalhos da literatura relacionados a este. Na seção 3 é apresentada a metodologia que está sendo adotada no desenvolvimento desta pesquisa (ainda em andamento), bem como os recursos e ferramentas linguísticos computacionais usados. Na seção 4 são apresentados alguns resultados preliminares. Por fim, a seção 5 apresenta algumas conclusões do trabalho.

2. Trabalhos Relacionados

Na literatura a análise de sentimentos (AS) tem sido aplicada a documentos, sentenças e aspectos [Medhat et al. 2014, Schouten and Frasincar 2016, Pereira 2021]. No nível de documentos, o sentimento é atribuído ao documento como um todo. No nível de sentenças, o sentimento é atribuído a cada sentença de um documento, enquanto no nível de aspectos os sentimentos relacionados a aspectos/atributos específicos das entidades mencionadas no texto (alvos de opinião) são identificados e posteriormente classificados de acordo com a opinião ou emoção. Por exemplo, em um comentário sobre um produto qualquer (por exemplo, *smartphone*) o produto em si costuma ser o alvo de opinião, enquanto os atributos/características relacionadas ao produto são os aspectos (como preço, qualidade da câmera e duração da bateria). Neste artigo, o foco de interesse é a AS orientada aos alvos de opinião em textos. Ou seja, uma tarefa que se aproxima da AS em nível sentencial, porém é mais específica e desafiadora do que a análise geral do sentimento expresso em uma sentença. Por outro lado, é uma tarefa menos refinada do que a análise de sentimentos baseada em aspectos.

Uma abordagem clássica de AS usada na literatura baseia-se no uso de léxicos de sentimentos, ou seja, dicionários que contêm palavras anotadas com suas respectivas polaridades. Esse tipo de abordagem utiliza a contagem e/ou a soma das polaridades das palavras presentes no texto para determinar sua polaridade geral (vide [Taboada et al. 2011, Liu and Zhang 2012, Costa and Pardo 2020]). Além das abordagens baseadas em léxicos, técnicas baseadas em aprendizado de máquina também têm se destacado em análise de sentimentos, incluindo tanto as técnicas tradicionais de aprendizado supervisionado e não supervisionado (por exemplo, [França and Oliveira 2014, Cristiani et al. 2020, Capellaro and Caseli 2021, Jain et al. 2021]) como técnicas mais sofisticadas baseadas em aprendizado profundo (*deep learning*) (por exemplo, [Zhang et al. 2018, Souza and Oliveira e Souza Filho 2022]). Abordagens híbridas combinando várias técnicas também são comuns [Appel et al. 2016].

Em [Carvalho et al. 2017], por exemplo, os autores compararam três diferentes classificadores (Naive Bayes, SVM e MaxEnt) e três métodos de seleção de atributos na classificação de polaridade em parágrafos de textos relacionados às eleições de 2014 no Brasil. O melhor classificador (MaxEnt) obteve uma acurácia em torno de 85%. Já o modelo de [Capellaro and Caseli 2021], baseado no BERT pré-treinado para o português, alcançou um *F1-score* de 96,6% na classificação de polaridade associada ao sentimento geral de *tweets* relacionados às eleições de 2018 no Brasil.

Se por um lado as abordagens baseadas em aprendizado de máquina podem levar a resultados bastante precisos na classificação de polaridade de sentimentos de textos para

o domínio no qual ele foi treinado, por outro lado esses modelos costumam ter um desempenho muito ruim quando aplicados a outros domínios [Taboada et al. 2011]. Nesse sentido, as abordagens baseadas em léxicos de sentimentos e as abordagens híbridas que combinam esse tipo de recurso com métodos de aprendizado de máquina são interessantes e podem resultar em modelos menos dependentes de domínio. Nesta pesquisa, ainda em andamento, iniciamos a investigação com o uso de abordagens baseadas em léxicos, para no futuro tratar a tarefa de análise de polaridade como um problema de aprendizado de máquina.

3. Metodologia de Desenvolvimento

A metodologia de desenvolvimento deste trabalho é baseada em corpus de comentários do domínio político e no uso de léxicos de sentimentos. Mais especificamente, como corpus de trabalho foi escolhido o SentiCorpus-PT [Carvalho et al. 2011], composto por comentários sobre debates televisivos referentes às eleições de 2009 do Parlamento Português. O corpus contém 1.082 comentários, totalizando 3.868 sentenças. Cada sentença no corpus pode ter diferentes alvos de opinião. Os alvos de opinião são entidades humanas, nomeadamente políticos, organizações políticas (geralmente utilizadas para se referir aos seus membros), personalidades da mídia (por exemplo, jornalistas) ou usuários (comentadores). Cada sentença dispõe de anotações sobre cada alvo de opinião mencionado na sentença e a polaridade relacionada a cada alvo. A polaridade varia de -2 (o valor negativo mais forte) até 2 (o valor positivo mais forte). 94,3% das sentenças possuem pelo menos um alvo anotado, sendo que a maioria delas (79%) tem exatamente um alvo de opinião.

Como léxicos de sentimentos tem sido utilizado o LIWC [Balage Filho et al. 2013] e o SentiLex-PT [Carvalho and Silva 2015]. O LIWC é um léxico geral do português composto por 127.149 instancias organizadas em categorias. As categorias *posemo* e *negemo* indicam polaridade positiva e negativa, respectivamente. O SentiLex-PT, por sua vez, é um léxico de sentimentos sobre entidades humanas. Ele é composto por 7.014 lemas e 82.347 formas flexionadas organizadas em adjetivos, substantivos, verbos e expressões idiomáticas.

Para o pré-processamento do corpus foi utilizada a biblioteca Python spaCy¹ com o modelo português "pt_core_news_lg". O pré-processamento consistiu nas seguintes etapas: (i) tokenização, (ii) lematização, (iii) extração de *PoS - Part of Speech* e (iv) análise sintática de dependência.

Após o pré-processamento do corpus, um primeiro estudo foi realizado para atribuir a polaridade de sentimento (positiva, negativa ou neutra) para cada sentença em relação a cada alvo de opinião presente na mesma. Mais especificamente, a estratégia usada consistiu em atribuir ao alvo de opinião do comentário a polaridade resultante da soma das polaridades das palavras presentes no léxico (LIWC ou SentiLex-PT). Para cada léxico, duas diferentes estratégias foram usadas: a primeira considera apenas as palavras associadas ao alvo do comentário, via dependência sintática e a segunda que considera o comentário como um todo.

¹<https://spacy.io/models/pt> (acesso em: 02/07/2023).

4. Resultados

Table 1. Resultados obtidos por cada estratégia.

Estratégia	Precisão	Cobertura	Medida-F
SL	45,1%	31,2%	36,9%
LW	36,6%	26,8%	30,9%
SL+LW	39,8%	34,0%	36,7%
SL-DEP	55,1%	5,5%	10,1%
LW-DEP	44,5%	5,1%	9,1%
SL+LW-DEP	49,5%	7,5%	13,0%

Em um experimento preliminar, cada estratégia de classificação de polaridade foi avaliada em termos de Precisão, Cobertura e Medida-F. A Tabela 1 resume os resultados obtidos com cada estratégia. Em termos de Precisão, o melhor valor (55,1%) foi obtido pela estratégia SL-DEP que se baseia no uso do SentiLex e nas relações de dependência sintática para associar cada alvo de opinião a cada palavra de sentimento identificada pelo léxico. Porém, as três estratégias que usam dependência sintática apresentaram os piores valores de Cobertura e, conseqüentemente, os piores desempenhos globais medidos em termos de Medida-F. Uma possível explicação para isso pode estar relacionada ao fato de que vários comentários no corpus não apresentam a estrutura esperada de sujeito-verbo-objeto, o que pode ter impactado na qualidade da análise de dependência gerada, fazendo com que a palavra com polaridade não fosse corretamente associada ao alvo do comentário. O melhor desempenho em termos de Cobertura (34%) foi obtido ao combinar o uso do SentiLex e do LIWC em uma mesma estratégia. No que se refere ao desempenho global, a estratégia baseada apenas no uso do SentiLex e a estratégia que combinou o uso dos dois léxicos (SL+LW) obtiveram os melhores desempenhos, ou seja, 36,9% e 36,7%, respectivamente. Contudo, testes estatísticos precisam ser realizados a fim de confirmar se essa pequena diferença entre essas duas abordagens é estatisticamente significativa.

5. Conclusões

Este trabalho descreveu um estudo preliminar com o objetivo de avaliar o uso de léxicos de sentimentos na classificação de sentimento (polaridade) relacionado a cada alvo de opinião em comentários sobre debate político em português. A tarefa de atribuir polaridade orientada aos alvos de opinião é mais desafiadora do que a classificação de sentimento geral de um comentário.

Como próximos passos deste trabalho pretende-se investigar outros léxicos disponíveis para o português e o uso de heurísticas definidas a partir de corpus que permitam associar palavras que expressam sentimento a cada alvo de opinião no texto. Pretende-se, ainda, combinar o uso de léxicos de sentimentos com técnicas de aprendizado de máquina.

References

- Appel, O., Chiclana, F., Carter, J., and Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124. New Avenues in Knowledge Bases for Natural Language Processing.

- 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.
- Capellaro, L. and Caseli, H. M. (2021). Análise de polaridade e de tópicos em tweets no domínio da política no Brasil. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 47–55, Porto Alegre, RS, Brasil. SBC.
- Carvalho, C. M. A., Nagano, H., and Barros, A. K. (2017). A comparative study for sentiment analysis on election Brazilian news. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 103–111, Uberlândia, Brazil. Sociedade Brasileira de Computação.
- Carvalho, P., Sarmiento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 564–568, Portland, Oregon, USA. Association for Computational Linguistics.
- Carvalho, P. and Silva, M. (2015). SentiLex-PT: Principais características e potencialidades. *Linguística, Informática e Tradução: Mundos que se Cruzam, Oslo Studies in Language*, 7(1):425–438.
- Chaves, M., Freitas, L., Souza, M., and Vieira, R. (2012). Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. volume 7337, pages 296–301.
- 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.
- Cristiani, A., Lieira, D., and Camargo, H. (2020). A sentiment analysis of brazilian elections tweets. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 153–160, Porto Alegre, RS, Brasil. SBC.
- França, T. and Oliveira, J. (2014). Análise de sentimento de tweets relacionados aos protestos que ocorreram no Brasil entre junho e agosto de 2013. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 128–139, Porto Alegre, RS, Brasil. SBC.
- Jain, P. K., Pamula, R., and Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review*, 41:100413.
- Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *Mining Text Data*, pages 415–463.
- Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.
- Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

- Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.
- Souza, F. D. and Oliveira e Souza Filho, J. B. (2022). BERT for sentiment analysis: Pre-trained and fine-tuned alternatives. 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 209–218, Cham. Springer International Publishing.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.
- Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8.

Aryon: um aplicativo Shiny para documentação e análise de línguas indígenas brasileiras

Mateus Zaparoli¹, Katiuska Rowe², Magnun Rochel Madruga²

¹Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

²Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

{mateuszaparoli7,katyfigueiredo0111}@gmail.com, magnun@ufmg.br

Abstract. *In order to gather scientific efforts for the documentation of Brazilian languages, this work presents partial results of the development of an R Shiny web application for documentation and exploratory analysis of indigenous languages. The objective is to concentrate lexicographic descriptions of different languages by different years and researchers in a single platform. It aims at making the platform a reliable source for querying quantitative information about different language levels. In the current version, the application provides vocabularies of two languages, namely: a) **Krenak**, a language spoken in the East of Minas Gerais, belonging to the Botocudo family of the Macro-Jê trunk; and b) **Terena**, a language of the Aruak family spoken in Mato Grosso do Sul and in the west of the State of São Paulo. The vocabularies available in the application are Bruno Rudolph [Rudolph 1909], for Krenak, and Denise Silva [Silva 2013], for Terena language.*

Resumo. *No sentido de reunir esforços científicos para documentação das línguas brasileiras, este trabalho apresenta resultados parciais do desenvolvimento de uma aplicação web R Shiny para a documentação e análise exploratória de línguas indígenas. O objetivo é concentrar em uma plataforma descrições lexicográficas de diversas línguas por diferentes épocas e pesquisadores. Além disso, objetiva-se também tornar a plataforma uma fonte confiável para consulta de informações quantitativas sobre os diferentes níveis linguísticos. Na versão atual, a aplicação disponibiliza vocabulários de duas línguas, a saber: a) **Krenak**, língua falada no leste de Minas Gerais, pertencente à família Botocudo do tronco Macro-Jê; e b) **Terena**, língua da família Aruak falada no Mato Grosso do Sul e no oeste do Estado de São Paulo. Os vocabulários disponíveis no aplicativo são os de Bruno Rudolph [Rudolph 1909], para o Krenak, e o de Denise Silva [Silva 2013], para língua Terena.*

1. Introdução

A Organização das Nações Unidas (ONU) e UNESCO afirmam que pelo menos 50% das línguas faladas de hoje estarão extintas ou seriamente ameaçadas até 2100. Por outro lado, as previsões mais pessimistas – e talvez realistas – afirmam que 90-95% serão extintas ou estarão seriamente ameaçadas até o final do século XXI. Nesse contexto, as

línguas indígenas brasileiras são parte das línguas que apresentam um alto grau de perigo de extinção nas próximas décadas, como consequência da colonização do Brasil e da expansão do português para a formação da identidade nacional. Conforme a Unesco, a humanidade pode resultar em apenas 300-600 línguas orais não ameaçadas até o final deste século [Moseley 2010], o que representaria somente algo entre 4% e 8% da diversidade linguística verificada atualmente.

O Brasil possui uma enorme diversidade linguística e étnica em seu território. As línguas indígenas brasileiras pertencem a diferentes grupos linguísticos, que se subdividem em importantes famílias relacionadas ou a ramos de línguas, tais como: Macro-Jê, Tupi, Caribe, Pano, Tucano, Aruaque, Katukina, Maku, Nambikwara, Chapacura, Yanomami, Mura-Pirahã e Guaicuru. Os troncos linguísticos Macro-Jê e Tupi são exclusivos do Brasil, enquanto as demais famílias de línguas estendem-se a países vizinhos.

Acredita-se que o Brasil tenha aproximadamente 150 línguas atualmente faladas, embora não haja concordância quanto ao número exato. Em um trabalho minucioso sobre essa questão, [D'Angelis 2019] afirma que, de forma otimista, pode-se dizer que o Brasil apresenta atualmente em torno de 100 línguas. O cenário é preocupante, sobretudo no se refere à documentação e revitalização dessas línguas ainda vivas. Conforme [Moseley 2010], 13% têm uma descrição gramatical parca da língua, 38% possuem uma descrição avançada, 29% têm descrição científica e 19% delas apresentam descrição científica insignificante. Embora o trabalho de descrição tenha aumentado no Brasil em função dos esforços das universidades brasileiras e outras organizações internacionais, as documentações sobre as línguas são esparsas, de difícil acesso em bibliotecas do Brasil ou mesmo do mundo. Os materiais, sobretudo os lexicográficos, são raros ou carecem de digitalização e tratamento adequado.

No sentido de reunir esforços científicos para documentação das línguas brasileiras, este trabalho apresenta resultados parciais do desenvolvimento de uma aplicação web R Shiny para a documentação e análise exploratória de línguas indígenas. O objetivo é concentrar em uma plataforma descrições lexicográficas de diversas línguas por diferentes épocas e pesquisadores. Além disso, objetiva-se também tornar a plataforma uma fonte confiável para consulta de informações quantitativas sobre os diferentes níveis linguísticos. Sobre o projeto, o nome **Aryon** homenageia o grande pesquisador brasileiro *Aryon Dall'Igna Rodrigues*, pioneiro nos estudos das línguas indígenas da América do Sul. Seu livro *Línguas brasileiras: para o conhecimento das línguas indígenas* [Rodrigues 1994] é considerado um dos cem livros do século, portanto, um clássico do pensamento científico brasileiro.

2. Aryon: uma aplicação web R Shiny

O **Projeto Aryon** é uma aplicação web implementada na linguagem R que utiliza recursos do pacote Shiny (cf. [R Core Team 2021] [Chang et al. 2023]), cujo intuito é oferecer às comunidades indígenas e científicas uma ferramenta voltada à consulta de vocabulários e à visualização de gráficos e estatísticas sobre estruturas que perpassam os níveis fonético, fonológico, morfológico, sintático e semântico das línguas brasileiras.

Na versão atual, a aplicação disponibiliza vocabulários de duas línguas, a saber: a) **Krenak**, língua falada no leste de Minas Gerais, pertencente à família Botocudo do tronco Macro-Jê; e b) **Terena**, língua da família Aruak falada no Mato Grosso do Sul e no

oeste do Estado de São Paulo. Os vocabulários disponíveis no aplicativo são os de Bruno Rudolph [Rudolph 1909], para o Krenak, e o de Denise Silva [Silva 2013], para língua Terena.

The screenshot shows the ARYON application interface. At the top, there's a header with the ARYON logo and a search bar. Below the header, there's a section for selecting the language and author. The main part of the interface is a table displaying search results. The table has columns for 'Lema', 'Terena', 'Português', 'Português 2', 'Autor', and 'Ano'. The results are filtered for the word 'atapi' in Terena. The table shows several entries with their corresponding Portuguese translations and the author 'Denise Silva' from the year '2012'.

Lema	Terena	Português	Português 2	Autor	Ano
A	atapi	quarta		Denise Silva	2012
A	atapi	dar conta		Denise Silva	2012
A	atapi	apresentado para alguém que está sendo explicado por outro, terceiro		Denise Silva	2012
A	atapi	fruto do mal, parecido com o português		Denise Silva	2012
A	atapi	semente	reção, base de amarrado, pedra de estique	Denise Silva	2012
A	atapi	coque, gentio masculino		Denise Silva	2012
A	atapi	atapi		Denise Silva	2012
A	atapi	atapi		Denise Silva	2012
A	atapi	atapi		Denise Silva	2012
A	atapi	atapi		Denise Silva	2012
A	atapi	atapi		Denise Silva	2012

Figura 1. Apresentação da Estrutura de Vocabulários no Aryon

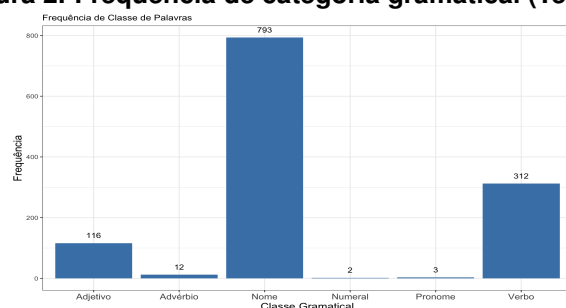
A Figura 1 mostra uma das abas do aplicativo em que se disponibilizam os vocabulários. Nessa aba, o usuário pode consultar pela palavra na língua indígena, português ou em outra língua em que ela tenha sido originalmente descrita. No caso do Krenak, por exemplo, esta língua foi descrita alemão, língua materna do autor. A tradução para o português foi dada pela equipe do projeto. É possível, ainda, buscar por autor, ano ou família linguística. Na seção a seguir, apresentamos um exemplo de análise exploratória possível de ser realizada no aplicativo.

3. Explorando o léxico Terena (Aruak)

Através da disponibilização do vocabulário, faz-se uma anotação linguística da categoria gramatical das palavras disponíveis. Com essa informação, pode-se explorar o funcionamento do léxico, sua composição em categorias gramaticais, o tamanho de cada classe e o número de palavras das classes aberta e fechada.

Na Figura 2, apresentamos a frequência das classes de palavras da língua Terena, conforme análise gramatical disponibilizada em [Silva 2013].

Figura 2. Frequência de categoria gramatical (Terena)



O vocabulário de Silva disponibiliza centenas de frases na língua e suas respectivas traduções em português. Com isso, é possível analisar, de forma imediata, a morfologia e sintaxe do Terena. O léxico da autora disponibilizado no **Aryon** conta com 312 entradas e 1245 ocorrências, numa razão tipo/ocorrência de 0.25.

4. Considerações Finais

O trabalho apresentado tem como objetivo criar uma ferramenta computacional que seja capaz de congrega informações confiáveis para pesquisas em Linguística, mas sobretudo uma ferramenta capaz de ser utilizada pelas comunidades indígenas brasileiras em suas escolas, computadores e celulares. Como está em fase inicial, o projeto está em fase de aprimoramento da interface de consulta pelos usuários.

Referências

- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2023). *shiny: Web Application Framework for R*. R package version 1.7.4.9002.
- D'Angelis, W. d. R. (2019). Línguas indígenas no brasil: quantas eram, quantas são, quantas serão. *Revitalização de línguas indígenas: o que é*, pages 13–26.
- Moseley, C. (2010). *Atlas of the World's Languages in Danger*. Unesco.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rodrigues, A. D. (1994). *Línguas brasileiras: para o conhecimento das línguas indígenas*, volume 11. Edições Loyola.
- Rudolph, B. (1909). *Wörterbuch der Botokudensprache*. Thaden.
- Silva, D. (2013). Estudo lexicográfico da língua terena: proposta de um dicionário terena-português.

Desambiguação dos termos do Atlas Linguístico do Brasil através da OpenWordnet-PT-ALiB

Augusto Sampaio Barreto, Daniela Barreiro Claro

¹FORMAS Research Group

Instituto de Computação, Universidade Federal da Bahia
Salvador - Bahia - Brazil

{augusto.barreto, dclaro}@ufba.br

Abstract. *This work describes the disambiguation of terms from the Linguistic Atlas of Brazil (ALiB) via OpenWN-PT-ALiB through a Twitter corpus. The study presents two main contributions: the incorporation of some ALiB terms in OpenWordNet-PT (OpenWN-PT) and the development of a disambiguation method using Word Embeddings and the Soft Cosine Measure (SCM). The proposed method uses Word Embeddings to represent the words in a vector space and calculates the SCM between the context of the tweets and the possible synsets of OpenWN-PT-ALiB for disambiguation. Results demonstrate the effectiveness of the method, with higher disambiguation rates even in the context of Twitter.*

Resumo. *Este trabalho descreve a desambiguação de termos do Atlas Linguístico do Brasil (ALiB) via OpenWN-PT-ALiB através de um corpus do Twitter. O estudo apresenta duas principais contribuições: a incorporação de alguns termos do ALiB na OpenWordNet-PT (OpenWN-PT) e o desenvolvimento de um método de desambiguação utilizando Word Embeddings e a Soft Cosine Measure (SCM). O método proposto utiliza Word Embeddings para representar as palavras em um espaço vetorial e calcula a SCM entre o contexto dos tweets e os possíveis synsets da OpenWN-PT-ALiB para a desambiguação. Os Resultados demonstram a eficácia do método, com taxas de desambiguação superiores mesmo no contexto do Twitter.*

1. Introdução

O ALiB começou em 2001, encerrando seus inquéritos na década passada. Os termos coletados ao longo de duas décadas de inquéritos suscitam a análise de sua presença nas redes sociais devido à disseminação dessas plataformas e à quantidade significativa de postagens em linguagem escrita. A difusão das redes sociais e o grande volume de postagens destacam a importância de investigar a presença dos termos coletados pelo ALiB nessas plataformas.

Analisar a vitalidade linguística dos termos do ALiB nas redes sociais exige automação devido ao volume de tweets. Muitos termos do ALiB são ambíguos e necessitam de desambiguação conforme o contexto do Atlas. Por exemplo, *bala* em tweets geralmente se refere a munição, enquanto no ALiB pode significar um tipo de doce.

Assim, o presente trabalho evidencia duas principais contribuições com o intuito de automatizar a desambiguação dos termos do ALiB. A primeira contribuição se refere

à OpenWordNet-PT (OpenWN-PT), à qual foram adicionados os *synsets* de alguns dos termos do ALiB. A segunda contribuição se refere ao método de desambiguação dos termos do ALiB para usufruir da OpenWordNet-PT-ALiB (OpenWN-PT-ALiB) que foi gerada. O método proposto se baseia no uso de *Word Embeddings* para encapsular as palavras no espaço vetorial, e no cálculo da *Soft Cosine Measure* (SCM) entre o contexto dos tweets com a palavra a ser desambiguada e os possíveis *synsets* existentes para essa palavra na OpenWordNet-PT-ALiB.

2. OpenWordNetPt-ALiB

A OpenWordnet-PT [de Paiva et al. 2012], abreviado como OpenWN-PT, surgiu com base no projeto da WordNet [Fellbaum 1998], de modo a atender à demanda da comunidade científica por ferramentas disponíveis em língua portuguesa de forma gratuita, acessível online e disponível para download e uso offline. Ela mapeia os *synsets* da língua inglesa para os correspondentes em português.

Com base no arquivo original em formato RDF da OpenWN-PT, alguns dos *synsets* do ALiB foram incorporados diretamente em seu arquivo fonte no formato RDF, sendo esta uma das contribuições científicas deste trabalho. A análise lexical incluída na nova versão, a OpenWN-PT-ALiB, foi baseada no questionário semântico-lexical obtido do Projeto ALiB. Variantes lexicais foram fornecidas por informantes brasileiros localizados em todo o território nacional. Esses dados foram publicados nos dois volumes do Atlas Linguístico do Brasil [Cardoso and Mota 2014]

Com o intuito de analisar se os resultados do ALiB seriam retornados dada uma consulta a um *synset*, o termo *Goleiro* foi utilizado.

A Figura 1 apresenta o resultado da consulta ao *synset* do termo *Goleiro* no arquivo original da OpenWN-PT.

```
WORD: goleiro  
WORD_SENSES: ['goleiro', 'guarda-redes']
```

Figure 1. Consulta aos *synsets* da palavra “goleiro” na OpenWN-PT

Já a Figura 2 apresenta o resultado da consulta ao *synset* do termo *Goleiro* na OpenWN-PT-ALiB, após adição do novo sentido “sutiã”

```
WORD: goleiro  
WORD_SENSES: ['sutiã', 'goleiro', 'guarda-redes']
```

Figure 2. Consulta aos *synsets* da palavra “goleiro” OpenWN-PT-ALiB

3. Modelos de Linguagem

Diversos modelos de linguagens foram propostos [Bengio et al. 2003], com o objetivo de representar o espaço semântico ideal de palavras em um espaço vetorial contínuo com valor real. As representações distribuídas dos termos em um espaço vetorial (*word embeddings*) ajudam os algoritmos de aprendizado a obter melhores desempenhos em tarefas de processamento de linguagem natural, agrupando palavras similares [Mikolov et al. 2013]

O NILC-Embeddings [Hartmann et al. 2017] é um repositório público que foi desenvolvido com o objetivo de compartilhar *word embeddings* gerados para a Língua Portuguesa. Ele contribui para tornar acessível recursos vetoriais a serem utilizados em tarefas de Processamento da Linguagem Natural e Aprendizado de Máquina para língua portuguesa.

Neste trabalho, os vetores de embeddings pré-treinados disponibilizados no projeto NILC-Embeddings foram utilizados para a tarefa de desambiguação de sentido de palavras, do inglês Word Sense Disambiguation (WSD).

4. Desambiguação automática dos termos do ALiB

Em termos gerais, a tarefa de *word sense disambiguation* envolve a associação de uma determinada palavra em um texto com uma definição ou significado (sentido) que é distinguível de outros significados potencialmente atribuíveis a essa palavra. A tarefa, portanto, necessariamente envolve duas etapas: (1) a determinação de todos os diferentes sentidos para cada palavra relevantes (pelo menos) para o texto em consideração; e (2) um meio para atribuir cada ocorrência de uma palavra ao sentido apropriado [Ide and Véronis 1998].

O Soft Cosine Measure (SCM), uma extensão da medida de similaridade do cosseno, oferece a capacidade de avaliar a similaridade entre dois documentos, mesmo quando não compartilham palavras em comum. Inicialmente proposto por Mikolov et al. (2013), esse método emprega uma medida de similaridade entre palavras, que é obtida por meio de operações vetoriais entre as *word embeddings* individuais das palavras. No contexto desse estudo, a SCM foi empregada para calcular a similaridade entre o contexto de uma palavra sujeita a desambiguação e a sua definição na base OpenWN-PT/ALiB.

Com o intuito de desambiguar os termos do ALiB, um método foi proposto, como mostrado na Figura 3.

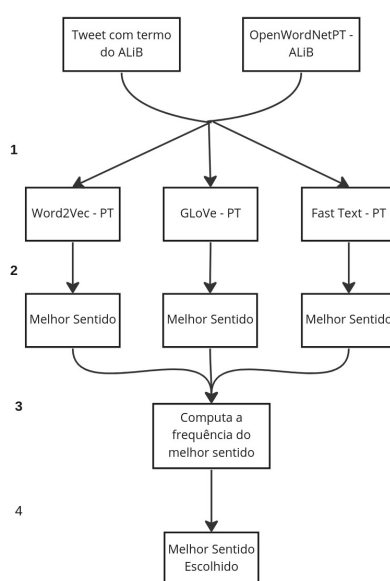


Figure 3. Representação arquitetural do método de desambiguação utilizado neste trabalho.

Nesta arquitetura, é importante evidenciar quatro macros etapas:

1. Cálculo da SCM pelos modelos de word embeddings CBOW de 600 dimensões, utilizando o contexto dos tweets minerados com a palavra a ser desambiguada, e os *synsets* desta palavra na OpenWN-PT/ALiB
2. Escolha do melhor sentido, com base na maior similaridade encontrada com base no cálculo da SCM.
3. Totaliza o melhor sentido escolhido por cada modelo.
4. Escolhe o melhor sentido para desambiguação, com base no que foi escolhido pela maioria dos modelos.

5. Experimentos e Resultados

O corpus, denominado TweetALiB/PT, composto de tweets coletados com o uso da biblioteca Tweepy foi utilizado.

Os experimentos foram realizados para identificar a vitalidade dos termos ALiB no Twitter e comparar os métodos de Word Sense Disambiguation (WSD) para a língua portuguesa em Tweets. Para comparar os resultados, foi escolhida uma amostra aleatória de 100 tweets para cada termo com sentido a ser desambiguado.

Como não existe um dataset rotulado com o sentido original dos tweets minerados para o ALiB, foi necessária realizar uma conferência manual na tarefa de desambiguação dos sentidos dos tweets. Para isso, 100 tweets aleatórios foram selecionados e o procedimento de desambiguação foi realizado manualmente, comparado com o sentido encontrado pelo método proposto.

Um importante resultado a se destacar é que, através do método proposto, a tarefa de desambiguação automática pode ser realizada em maior escala. Utilizando o SCM, a desambiguação nos experimentos atingiu valores superiores a 25 %, chegando a 55 % para a palavra “goleiro”.

Ressalta-se que o Twitter é uma plataforma em que a linguagem empregada é majoritariamente informal e que se resume a poucas palavras de contexto, dificultando a tarefa de desambiguação até mesmo para um humano.

6. Conclusão e Trabalhos Futuros

O presente trabalho descreveu o método referente à incorporação de alguns dos termos do ALiB na OpenWordnetPT-ALiB. Os resultados obtidos, evidenciam que ainda existem desafios na tarefa de desambiguação utilizando o método proposto em compreender as nuances semânticas e realizar associações relevantes em textos curtos e informais, característicos da plataforma. Como trabalhos futuros, pode-se explorar a investigação de técnicas de pré-processamento de texto específicas para lidar com a linguagem peculiar do Twitter.

Agradecimentos

O presente trabalho conta com o apoio da CAPES-Brasil - Código de Financiamento 001 e da FAPESB - Projeto TIC.

References

- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.
- Cardoso, S. and Mota, J. (2014). *Atlas Linguístico do Brasil*. Addison-Wesley Longman Publishing Co., Inc.
- de Paiva, V., Rademaker, A., and de Melo, G. (2012). Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India. The COLING 2012 Organizing Committee. Published also as Techreport <http://hdl.handle.net/10438/10274>.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.
- Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil. SBC.
- Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

TransAlign: tradução e alinhamento de corpora para a língua portuguesa

Alan Rios Melo¹, Daniela Barreiro Claro¹

¹FORMAS Research Group
Instituto de Computação, Universidade Federal da Bahia
Salvador - Bahia - Brazil

{alan.rios, dclaro}@ufba.br

Abstract. *In this paper, we introduce TransAlign, an innovative framework to enhance Open Information Extraction (OpenIE) in underrepresented languages, such as Portuguese, by leveraging data from resource-rich languages. Utilizing specific grammatical rules and high-quality translation models, we adapted LSOIE, a large-scale dataset, for Portuguese. This approach generated 21.161 high-quality triples for OpenIE in Portuguese. The resulting dataset enabled the training of a new model that improved F1 scores by 50% over existing systems for Portuguese*

Resumo. *Neste artigo, apresentamos o TransAlign, uma estrutura inovadora para ampliar a Extração Aberta de Informações (OpenIE) em idiomas sub-representados, como o português, usando dados de idiomas ricos em recursos. Utilizando regras gramaticais específicas e modelos de tradução de alta qualidade, adaptamos o LSOIE, um conjunto de dados de grande escala, para o português. Essa abordagem gerou 21.161 triplas de alta qualidade para OpenIE em português. O conjunto de dados resultante possibilitou o treinamento de um novo modelo que melhorou em 50% os escores F1 dos sistemas existentes para o português*

1. Introdução

A Extração Aberta de Informações (OpenIE) é um processo essencial no processamento de linguagem natural (PLN) que extrai informações estruturadas de texto não estruturado [Banko et al. 2007, Etzioni et al. 2008]. Apesar dos avanços nos sistemas OpenIE para inglês [Angeli et al. 2015, Stanovsky et al. 2018], ainda há desafios para idiomas sub-representados [Akbik et al. 2019b]. Este artigo apresenta o TransAlign, um framework que utiliza traduções e regras de alinhamento específicas do idioma para criar conjuntos de dados OpenIE em idiomas sub-representados. Utilizamos o português como exemplo, gerando um novo conjunto de dados a partir do LSOIE [Solawetz and Larson 2019], resultando em melhorias significativas no desempenho do OpenIE em português. Nosso trabalho destaca o potencial de grandes conjuntos de dados e ferramentas de tradução para a pesquisa OpenIE em idiomas sub-representados. Este artigo está organizado como segue: a seção 2 descreve a validação dos dados, a seção 3 a construção do dataset e a seção 4 o modelo proposto, tendo a seção 5 os experimentos e resultados.

2. Verificador de correspondencia

Com o objetivo de assegurar o alinhamento correto de todos os dados utilizados no treinamento do modelo, foi desenvolvido um algoritmo de correspondência. Esse algoritmo busca identificar as correspondências dos elementos das triplas ('ARG0', 'REL', 'ARG1'), garantindo que cada elemento seja encontrado em conjunto com todos os tokens pertencentes à mesma sentença, sem qualquer interrupção de palavras externas às triplas. O algoritmo de correspondência possui duas abordagens: busca sequencial e busca não sequencial. Na busca sequencial, são selecionadas as sentenças que possuem exatamente a ordem (ARG0, REL, ARG1) dos elementos da tripla encontrados. Já na busca não sequencial, não há essa limitação, considerando todas as possíveis ordens de elementos como válidas. No método apresentado neste artigo, a abordagem sequencial foi descrita, devido à menor complexidade no treinamento dos modelos.

3. Construção do Dataset

A construção do conjunto de dados foi motivada pela escassez de dados em grande quantidade para a tarefa de OpenIE na língua portuguesa do Brasil. Tentativas anteriores de obtenção desses dados, utilizando diferentes técnicas, resultaram em dados inconsistentes e de qualidade mediana. Diante da necessidade de construir um conjunto de dados amplo e de melhor qualidade, a abordagem proposta traduziu e alinhou os conjuntos de dados do inglês para o português brasileiro.

3.1. TransAlign

Na primeira tentativa, o 'PTOIE', originado da tradução do 'SQuAD v2'[Rajpurkar et al. 2016], apresentou ruídos, levando a 7.344 extrações de média qualidade e 2.472 correspondidas de 130.000 instâncias iniciais.

O TransAlign, um conjunto de dados traduzido do inglês para o português, minimiza o ruído de conversão, embora a tradução possa introduzir algumas incompatibilidades. Utilizando a API do Google Tradutor e o GPT 3.5 para a tradução, foram obtidas 7.000 e 22.124 extrações de válidas, respectivamente, de um total de 49.566. Um desafio significativo é a preservação das características de anotação das sentenças durante a tradução, o que requer um processo de alinhamento cuidadoso. Este processo é composto por três etapas principais:

Busca de Combinações de Relações: O algoritmo busca todas as combinações possíveis de relações na sentença traduzida, que podem ocorrer entre o primeiro e último token.

Correspondência de Triplas e Frases: O algoritmo procura correspondências entre a tripla gerada e a frase inteira. Se uma correspondência é encontrada, a 'pos tag' é anotada com base na análise feita na sentença, utilizando o modelo treinado pelo spacy [Honnibal and Montani 2017].

Verificação de 'Pos Tags': O algoritmo percorre a 'postag' anotada da combinação de relação, a etapa é uma adaptação das regras apresentadas no reverb para a língua pt-br [Fader et al. 2011], verificando se a relação começa com um verbo e se o restante das tags da relação pertence a um advérbio, adjetivo, verbo, pronome ou substantivo. A estrutura da relação determina o último token necessário para a relação ser considerada válida.

Table 1. Estatísticas de Conversão TransAlign

	Original	TransAlign
LSOIE Train [Solawetz and Larson 2019]	49.566	15.006
LSOIE Test [Solawetz and Larson 2019]	10.783	3.282
LSOIE Dev [Solawetz and Larson 2019]	9.459	2.873
CARB [Bhardwaj et al. 2019]	3.497	715
S2 Train [Kolluru et al. 2022]	166.032	77.805
S2 Valid [Kolluru et al. 2022]	1.872	923
all	231.750	100.604

Se houver mais de uma combinação possível de relação, a maior delas é escolhida, pois a escolha da relação com maior quantidade de tokens melhora e complementa o contexto da extração. Após a seleção da relação, a extração é realinhada, com todos os tokens antes do primeiro token da relação considerados como argumento 0, e todos os tokens após o último token da relação considerados como argumento 1. Gerando assim, um conjunto de dados anotados OpenIE para o português do Brasil

- *Inglês:*

Sentença: *"English longbow was also used against the English by their Welsh neighbours."*

Tripla: *(English longbow – was also used – against the English by their Welsh neighbours)*

- *Português:*

Sentença: *"O arco longo inglês também foi usado contra os ingleses por seus vizinhos galeses."*

Tripla: *(O arco longo inglês – foi usado contra – os ingleses por seus vizinhos galeses)*

4. Modelo

O modelo utilizado para o treinamento foi desenvolvido com base na framework Flair-NLP [Akbik et al. 2019a], adotando uma metodologia de Rotulagem de Sequência. A arquitetura do modelo inicialmente incorpora um Word Embedding [Akbik et al. 2018], seguido por embeddings direcionais, tanto para frente quanto para trás [Akbik et al. 2018]. Esses três conjuntos de embeddings são então concatenados e direcionados para uma camada linear, que unifica e representa todos os codificadores. Após essa reprojeção, adicionamos duas camadas de Redes Neurais Recorrentes (RNN) para processamento subsequente. Na extremidade da arquitetura, posicionamos um classificador que conta com uma camada CRF (Campo Aleatório Condicional).

4.1. Fine-Tuning

Após a fase de treinamento, o modelo passa por um processo de ajuste fino, ou fine-tuning. Esse processo emprega o dataset Pud-200 [Cabral et al. 2022], uma seleção de extrações "silver" anotadas manualmente. Essa etapa de ajuste ocorre durante 20 épocas. Sendo este processo importante por se tratar de dados nobres que ajudam na precisão das extrações.

4.2. Datasets utilizados

O modelo TransAlign foi treinado com datasets traduzidos e alinhados, conforme apresentado na Tabela 1. O modelo PTOIE, por outro lado, foi treinado usando a técnica de conversão no conjunto de dados SQuAD v2. Todos os modelos foram validados usando o conjunto de dados PUD-100 [Cabral et al. 2022], que inclui 100 sentenças com extrações 'gold', que foram anotadas e validadas manualmente por três anotadores distintos. Apenas as extrações que receberam dois ou mais votos positivos foram selecionadas.

5. Experimentos e Resultados

Os modelos experimentais são comparados usando métodos estatísticos e qualitativos. As métricas estatísticas incluem precisão, recall e f1-score, gerados a partir da validação com o dataset Pud-100 na ferramenta CaRB [Bhardwaj et al. 2019]. A análise qualitativa envolve a revisão manual das triplas extraídas pelos modelos, proporcionando uma visão realista da qualidade e aplicabilidade prática dos modelos treinados.

O conjunto de dados exibiu um desempenho robusto no treinamento do modelo, resultando em triplas coerentes e contextualmente relevantes. O modelo alcançou métricas de benchmark, com f1-score de 0.3228 e 0.3766, precisão de 0.4137 e 0.4827, e cobertura de 0.2647 e 0.3088, no *identical match* e *lexical match*, respectivamente. Ao avaliar a capacidade do modelo de extrair fatos de um texto gerado pelo GPT-4 sobre furtos fictícios em uma cidade, o modelo demonstrou habilidade em realizar inferências precisas e extrair informações relevantes, o texto possui 249 palavras.

Algumas extrações:

”(Adrian, estava fora de, a cidade)”

”(Sua casa, estava, vazia)”

”(Clara uma senhora idosa, cuidava de, a casa)”

”(Clara, voltou de, sua caminhada noturna)”

”(A porta de a frente, estava, entreaberta)”

”(Clara, chamou, a polícia)”

”(Os policiais, chegaram em, 10 minutos)”

”(Os indícios, sugeriram que, os criminosos eram profissionais)”

”(O roubo, pareceu ter ocorrido, entre as 20h e as 20h30)”

6. Conclusão e Trabalhos Futuros

Este estudo produziu resultados significativos para a língua portuguesa na tarefa de EIA, demonstrando a viabilidade de gerar dados de alta qualidade e enriquecer idiomas sub-representados. A expectativa futura é superar desafios relacionados ao aumento da complexidade dos dados e à cobertura de diferentes estruturas, o que parece ser uma meta alcançável com base nos resultados atuais.

Agradecimentos

O presente trabalho conta com o apoio da CAPES-Brasil - Código de Financiamento 001 e da FAPESB - Projeto TIC.

References

- Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019a). Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.
- Akbik, A., Bergmann, T., and Vollgraf, R. (2019b). Pooled contextualized embeddings for named entity recognition. 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 724–728. Association for Computational Linguistics.
- Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.
- Angeli, G., Premkumar, M. J. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354. Association for Computational Linguistics.
- Banko, M., Cafarella, M., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial intelligence*, pages 2670–2676. University of Washington.
- Bhardwaj, S., Aggarwal, S., and Mausam, M. (2019). CaRB: A crowdsourced benchmark for open IE. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.
- Cabral, B., Souza, M., and Claro, D. B. (2022). Portnoie: A neural framework for open information extraction for the portuguese language. 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 243–255, Cham. Springer International Publishing.
- Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.
- Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. pages 1535–1545.
- Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Kolluru, K., Mohammed, M., Mittal, S., Chakrabarti, S., and ., M. (2022). Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.
- Solawetz, J. and Larson, S. (2019). LSOIE: A large-scale dataset for supervised open information extraction. *arXiv preprint arXiv:2101.11177*.
- Stanovsky, G., Michael, J., Zettlemoyer, L., and Dagan, I. (2018). Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895. Association for Computational Linguistics.

Desafios da tarefa de Extração de Informação Aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual

Beatriz Paixão Queiroz¹, Rerisson Cavalcante¹, Daniela Barreiro Claro²

¹FORMAS Research Group – Instituto de Letras, Universidade Federal da Bahia

²FORMAS Research Group – Instituto de Computação, Universidade Federal da Bahia
Salvador - Bahia - Brazil

{beatrizpaixao, rerisson.cavalcante, dclaro}@ufba.br

Abstract. *This work describes the methodology defined for the construction of a corpus, in Portuguese, manually annotated from an automated generation of a corpus for the Open Information Extraction task in Portuguese. Rules were defined for the extraction of triples in order to subsidize the generation of corpus in the creation of models based on machine learning. The results obtained were the generated corpus that has been used in the creation of algorithms for the EIA task.*

Resumo. *Este trabalho descreve a metodologia definida para a construção de um corpus, em português, anotado manualmente a partir de uma geração automatizada de um corpus para a tarefa de Extração de Informação Aberta em Português. Regras foram definidas para a extração de triplas com o objetivo de subsidiar a geração de corpus na criação de modelos baseado em aprendizado de máquinas. Os resultados obtidos foram o corpus gerado que vem sendo utilizado na criação de algoritmos para a tarefa de EIA.*

1. Introdução

A criação de corpora anotados é essencial para o treinamento e validação de recursos voltados para o Processamento de Linguagem Natural (PLN). Este trabalho traz uma abordagem metodológica para a construção de um corpus anotado, desde a geração automatizada de sentenças através da tradução até a geração manual para a tarefa de Extração de Informação Aberta (Open Information Extraction, OpenIE) [Banko et al. 2007] voltadas para o português brasileiro. A Extração de Informação Aberta consiste em gerar informação estruturada de textos não estruturados a qual normalmente é descrita através da tripla (arg0, rel, arg1) [Fader 2011]. Neste contexto, os vínculos semânticos entre entidades (pessoas, organizações, locais, datas etc) são mapeadas como relações.

As novas técnicas de aprendizagem de máquinas exigem grandes conjuntos de dados para que se possam treinar, validar e testar novos os algoritmos, fazendo assim avançar o estado da arte da OpenIE. A abordagem proposta consistiu da premissa de tradução de datasets em inglês para o português com o objetivo de fazer um alinhamento das extrações automatizadas. A partir das extrações inválidas, um novo corpus foi proposto com extrações manuais geradas por um linguista júnior e validadas por um linguista sênior, cujo objetivo é manter a alta qualidade do recurso criado. Esta metodologia foi composta por quatro etapas: (1) tradução do corpus; (2) análise das sentenças traduzidas,

para a extração manual de triplas; (3) anotação sintática manual dos componentes das triplas extraídas; (4) validação do corpus e a proposição de extrações adicionais.

Diversos trabalhos têm discutido metodologias para a criação de corpus grande, cujo objetivo é ser utilizado nas tarefas de PLN, especificamente em OpenIE [Stanovsky and Dagan 2016, Glauber et al. 2018]. Porém a principal limitação é a pequena quantidade de sentenças e de extrações, o que inviabiliza o uso de métodos baseados em aprendizado de máquinas. Assim, o principal objetivo deste trabalho é a criação de um corpus grande para ser utilizado em métodos baseados em redes neurais. As próximas seções detalham o conteúdo deste trabalho.

2. Etapa 1 : Tradução do SQuAD v2

O SQuAD v2 é um conjunto de dados desenvolvido para a tarefa de *Question Answering* (QA) em inglês e tem uma estrutura que permite sua conversão para a tarefa de *Open Information Extraction* (OpenIE). Neste processo de conversão, uma resposta do conjunto de dados original é interpretada como uma relação. Essa relação é então vinculada a um argumento primário (arg0), determinado por uma pergunta W (who, where, what, etc.), e é complementada com um segundo argumento (arg1), conforme adaptado dos autores em [Stanovsky and Dagan 2016].

As 7344 sentenças do SQuAD v2 foram traduzidas para o português por ferramentas automatizadas. Um subconjunto de 360 das sentenças traduzidas serviu de base para a tarefa de extração de construção do corpus manualmente anotado. Nesse processo, foram encontrados alguns erros de tradução nas sentenças iniciais. Com o intuito de validar esta primeira etapa, as sentenças foram revisadas e corrigidas por um linguista senior.

3. Etapa 2: Extração manual de triplas a partir das sentenças

Após a tradução dos dados do SQuAD v2, dois conjuntos de tarefas foram iniciadas, que resultaram em dois corpora de extrações (i) a validação manual das extrações geradas automaticamente pelo algoritmo, com a remoção das inválidas; (ii) a geração manual de novas extrações a partir de sentenças para as quais não foram registradas extrações automáticas coerentes ou seja, extrações inválidas.

A segunda dessas duas tarefas resultou no corpus descrito neste trabalho, com o acréscimo das etapas subsequentes: (a) anotação morfosintática manual dos constituintes das novas triplas propostas; (b) revisão das triplas propostas e das anotações. O objetivo principal foi a geração da maior quantidade possível de extrações válidas a partir da mesma sentença.

Para estruturação das regras, também foi utilizada uma das restrições sintáticas mencionadas pelos autores em [Fader 2011], na qual toda relação com múltiplas palavras pode iniciar com um verbo, terminar com uma preposição e consistir em uma sequência contígua de palavras.

3.1. Regras para extrações válidas

Regra 1: O argumento 0 (arg0) deve ser um sintagma nominal à esquerda do verbo, excluídos SN formados apenas por pronomes.

- Sentença (1): “Clésinger fez a máscara da mortuária de Chopin”.
- Extração: (Clésinger; fez; a máscara da mortuária de Chopin).

- Regra 2: O argumento 1 deve estar à direita do verbo e pode ser (i) um sintagma nominal, como no exemplo (1) acima; (ii) uma sequência formada por um sintagma nominal e outro(s) sintagma(s) complemento(s) ou adjunto(s), como no exemplo (2), respeitada a contiguidade; (iii) ou uma sentença, como no exemplo (3).
- Sentença (2): “*Peter Stuyvesant entregou Nova Amsterdã aos ingleses*”.
 - Extração: (Peter Stuyvesant; entregou; **Nova Amsterdã aos ingleses**)
 - Sentença (3): “*O sismólogo japonês Yuji Yagi disse que o terremoto ocorreu em duas etapas*”.
 - Extração: (O sismólogo japonês Yuji Yagi; disse; **que o terremoto ocorreu em duas etapas**).
- Regra 3: A relação deve conter pelo menos um verbo. Se houver uma preposição introduzindo o argumento 1, esta deve ser incluída na relação.
- Sentença: “*Léon Escudier escreveu sobre um recital de Chopin*”.
 - Extração: (Léon Escudier; **escreveu sobre**; um recital de Chopin).
- Regra 4: Se houver vários sintagmas complementos e adjuntos à direita do verbo, várias extrações foram feitas, com diferentes combinações dos elementos, respeitada a contiguidade.
- Sentença (4): “*Chopin visitou Berlim com um amigo*”.
 - Extração 1: (Chopin; visitou; **Berlin**)
 - Extração 2: (Chopin; visitou; **Berlin com um amigo**)
 - Extração 3: (Chopin; visitou Berlin com; **um amigo**)
 - Sentença (5): “*Peter Stuyvesant entregou Nova Amsterdã aos ingleses sem derramamento de sangue*”.
 - Extração 1: (Peter Stuyvesant; entregou; **Nova Amsterdã aos ingleses**)
 - Extração 2: (Peter Stuyvesant; entregou; **Nova Amsterdã aos ingleses sem derramamento de sangue**)
 - Extração 3: (Peter Stuyvesant; entregou Nova Amsterdã a; **os ingleses**)
 - Extração 4: (Peter Stuyvesant; entregou Nova Amsterdã a; **os ingleses sem derramamento de sangue**)
- Regra 5: A negação, outros advérbios e pronomes átonos imediatamente pré-verbais devem ser incluídos na relação.
- Sentença (6): “*A escola Theravada não inclui as escrituras Mahayava em seu cânon*”.
 - Extração: (a escola Theravada; **não inclui**; as escrituras Mahayava em seu cânon)
 - Sentença (7): “*John F. Shea se formou em 1908*”.
 - Extração: (John F. Shea; **se formou em**; 1908)
- Regra 6: A voz passiva permite diferentes extrações, com o particípio incluído na relação e no argumento 1.
- Sentença (8): “*A Batalha de Long Island foi travada em agosto de 1776*”
 - Extração: (A Batalha de Long Island”; **foi travada em**; agosto de 1776)
 - Sentença (9): “*Todas as rodovias em Wenchuan foram danificadas*”.
 - Extração: (Todas as rodovias em Wenchuan; foram; **danificadas**).

3.2. Regras para extrações inválidas

- Regra 1: Extrações cujos componentes fazem parte de orações diferentes são inválidas.
- Sentença (10): “*Portugal explorou o Oceano Atlântico, explorou a costa africana, colonizou áreas selecionadas da África, descobriu uma rota oriental*”

para a Índia (...)”.

- Extração inválida: (Portugal; colonizou; área selecionadas da África).
- Extração inválida (Portugal; descobriu; uma rota oriental para a Índia)
- Extração válida: (Portugal; explorou o Oceano Atlântico).

Regra 2: A extração é inválida se não há contiguidade entre os elementos internos que foram os argumentos da tripla.

- Sentença (11): “*Sassou venceu a seguinte eleição presidencial em julho de 2009*”.
- Extração inválida: (Sassou; **venceu em**; julho de 2009). - Extração válida (1): (Sassou; venceu; a seguinte eleição presidencial).
- Extração válida (2): (Sassou; venceu; a seguinte eleição presidencial em julho de 2009).
- Extração válida (3): (Sassou; venceu a seguinte eleição presidencial em; julho de 2009).

Regra 3: A extração é inválida se o argumento 0 não estiver à esquerda da relação e/ou se o argumento 2 não estiver à direita da relação

- Sentença (12): “*Existem, na Catalunha, Ilhas Baleares e Valência, padrões regionais*”.
- Extração inválida: (**padrões regionais**; existem em; a Catalunha, Ilhas Baleares e Valência).

Regra 4: A extração é inválida se algum dos argumentos for apenas um pronome.

- Sentença (13): “*Ela lhe deu um relógio de bolso*”.
- Extração inválida: (**ela**; lhe deu; um relógio de bolso).

4. Etapa 3: Anotação morfossintática manual dos componentes das triplas extraídas

Após a identificação manual das extrações de triplas, baseada no conjunto de regras definido, o processo de anotação morfossintática foi realizado utilizando etiquetas (tags) de Partes do Discurso (POS) [Alencar et al. 2018].

5. Etapa 4: Validação automática das extrações manuais

As extrações manuais foram feitas por uma linguista júnior e revistas por um linguista sênior, que também sugeriu novas extrações. Ao final do processo, foram obtidas 663 triplas a partir das 360 sentenças selecionadas dentre as 7344 sentenças traduzidas do SQuAD v2. Após isso, as extrações passaram por um algoritmo que busca correspondências nas respectivas sentenças, sendo que cada elemento da tripla deve ter seus tokens encontrados juntos e seguindo a ordem (arg0, rel, arg1). Das 663 triplas, 427 foram consideradas válidas pelo programa.

6. Conclusão e Trabalhos Futuros

O presente artigo apresentou uma metodologia com um conjunto de regras para auxiliar na tarefa de extração de informação aberta automatizada. Como trabalho futuro, este corpus gerado será validado através de um algoritmo de extração de triplas.

Agradecimentos

O presente trabalho conta com o apoio da CAPES-Brasil - Código de Financiamento 001 e da FAPESB - Projeto TIC.

References

- Alencar, L. F., Cuconato, B., and Rademaker, A. (2018). Morphobr: An open source large-coverage full-form lexicon for morphological analysis of portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.
- Banko, M., Cafarella, M., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial intelligence*, pages 2670–2676. University of Washington.
- Fader, Anthony, e. a. (2011). *Identifying Relations for Open Information Extraction*. Association for Computational Linguistics.
- Glauber, R., de Oliveira, L. S., Sena, C. F. L., Claro, D. B., and Souza, M. (2018). Challenges of an annotation task for open information extraction in portuguese. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonalo Oliveira, H., and Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, pages 66–76, Cham. Springer International Publishing.
- Stanovsky, G. and Dagan, I. (2016). Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Técnicas de sumarização de textos jurídicos para suporte à classificação de documentos de decisões judiciais

Hellen Harada¹, Fabíola Pereira¹, Alex Almeida², Daniela Freire³,
Márcio Dias⁴, Nádia Silva⁵, Pedro Andrade⁵, André Carvalho³

¹Faculdade de Computação
Universidade Federal de Uberlândia

²Faculdade de Tecnologia de Ourinhos

³Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

⁴Departamento de Ciência da Computação
Universidade Federal de Catalão

⁵Instituto de Informática
Universidade Federal de Goiás

{hellen.harada,fabiola.pereira}@ufu.br, alex.marino@fatecourinhos.edu.br,
{danielalfreire,andre}@usp.br, marciodias@ufcat.edu.br,
{andrdpedro,nadia.felix}@ufg.br

Resumo. *Acórdãos são documentos de texto que contêm decisões judiciais referentes a um certo processo jurídico. No contexto de um Tribunal de Justiça, os acórdãos possuem uma classificação por temas bem definida, que auxilia juristas na organização e agilidade de suas tarefas diárias. Devido ao alto volume diário de novos acórdãos produzidos, faz-se necessária a adoção de técnicas capazes de automatizar a classificação temática de um novo acórdão. Algoritmos de aprendizado de máquina supervisionado, para tarefas de classificação, não têm se saído bem diante de textos extensos, em português, com a linguagem usada no domínio jurídico. Este trabalho propõe a adoção de sumários de acórdãos para classificação temática. A hipótese levantada é que textos mais curtos, sumarizados, podem melhorar a classificação de tais documentos nos temas corretos. Este é um trabalho em andamento, que pretende desenvolver uma nova abordagem de classificação a partir de sumários. Resultados parciais indicam que algoritmos de sumarização melhoram a classificação de acórdãos.*

1. Introdução

Ao julgar um processo, o magistrado pode realizar despachos, decisões interlocutórias e sentenças. Os despachos, as decisões e sentenças são redigidos, datados e assinados pelos juízes, enquanto os acórdãos são feitos pelos desembargadores. Neste trabalho, os documentos judiciais abordados serão os acórdãos. Os acórdãos possuem uma classificação por temas do Superior Tribunal de Justiça (STJ), que são categorias baseadas nos fundamentos das decisões, bem como na legislação utilizada para embasar tais fundamentos. A estrutura de tematização é hierárquica, com a existência de subtemas, e dinâmica, com a

possibilidade de surgimento de novos temas. Seis temas gerais compõem o primeiro nível da árvore de hierarquia.

A classificação por temas é uma prática que auxilia os juristas na organização e agilidade do dia a dia, facilitando futuras decisões por similaridade entre os processos. O volume diário é de aproximadamente 100 novos acórdãos e, por isso, faz-se necessária a automatização do processo por meio de técnicas de Inteligência Artificial como um todo.

Os documentos dos acórdãos são, em geral, extensos e não possuem uma estrutura padrão. A classificação de acórdãos em temas pode ser modelada como um problema de aprendizado supervisionado. Entretanto, utilizar representações de texto que consideram o texto completo na construção dos modelos pode reduzir a eficácia da classificação, gerando indesejáveis instâncias de falsos positivos/negativos [Wang et al. 2021]. Sumarizar os textos em busca de seus segmentos mais representativos antes do processo de classificação pode ser uma promissora abordagem.

Dessa forma, a pergunta que se pretende responder neste trabalho é: *utilizar os acórdãos de maneira sumarizada pode aumentar a eficácia na classificação de tais documentos em relação aos seus temas?*

2. Fundamentação Teórica e Trabalhos Correlatos

Sumarização é a escrita de um texto mais curto comparado ao texto original e que permanece com a mesma ideia. Um resumo. Em termos de formação, sumários podem ser classificados como extrativos ou generativos.

Sumários extrativos são sumários compostos por partes inalteradas do texto original, de forma que o sumário seja composto das partes mais importantes do texto, sem haver modificações. Sumários generativos, também conhecidos como abstrativos são sumários feitos por meio da reescrita, havendo alteração dos seus trechos em comparação com o original, de forma que um pequeno texto passe toda a ideia principal do texto sem se preocupar com o modo de escrita original.

A sumarização automática de texto permite que os usuários compreendam e comparem rapidamente temas em determinados *corpora*. Tornou-se cada vez mais importante com o acúmulo crescente de documentos de texto em todos os campos [Wang et al. 2021].

Neste trabalho o foco é na sumarização extrativa por documento (monodocumento), ou seja, aquela que seleciona os segmentos mais importantes de um documento e os concatena para formar um sumário [El-Kassas et al. 2021].

Existem diversas técnicas bem estabelecidas para sumarização automática de textos, como por exemplo técnicas baseadas em grafos, em semântica e em centralidade de sentenças [El-Kassas et al. 2021]. Da mesma forma, existem muitas técnicas bem estabelecidas para classificação de textos, em especial textos jurídicos [Chen et al. 2022].

Entretanto, pouco ainda foi explorado acerca da classificação de textos com o auxílio de técnicas de sumarização. Em [Rahamat Basha et al. 2019] é proposto um novo método de seleção de características para o classificador KNN (*K-nearest neighbor*) resumindo os documentos de treinamento originais com base na medida de importância da sentença. A abordagem para sumarização de documento único usa duas medidas para similaridade de sentenças: a frequência dos termos em uma sentença e a similaridade dessa

sentença com outras sentenças.

Em [Jeong et al. 2016] os autores propõem um interessante *framework* que utiliza tanto informações de resumo quanto das categoriais. Um modelo de língua é utilizado para combinar distribuições de recursos em cada categoria e texto, e um modelo para classificação de texto faz as pontuações de importância de sentença estimadas a partir da sumarização de texto.

Em [Du et al. 2021] a sumarização é utilizada para construção de exemplos em uma abordagem de *text augmentation*, para resolver o problema de limitações de anotações.

Nenhum destes trabalhos possui foco em textos jurídicos.

3. Metodologia

A metodologia de desenvolvimento do trabalho está organizada em 4 etapas, descritas a seguir.

3.1. Desenvolvimento de abordagens de sumarização

O primeiro passo do trabalho foi o desenvolvimento de abordagens de sumarização existentes na literatura. O objetivo foi reunir 4 técnicas bem estabelecidas na literatura (KL-Soma, LexRank, LSA, Luhn) [El-Kassas et al. 2021], todas elas considerando a abordagem de sumarização extrativa monodocumento. Esta etapa foi importante para o entendimento do problema e percepção do impacto dos diferentes algoritmos.

3.2. Anotação do *corpus* por especialistas

Está sendo preparado um *corpus* para que especialistas realizem a anotação que, neste caso, consiste em gerar manualmente sumários em uma amostra de decisões judiciais, formando o *gold standard dataset*. Será utilizada a ferramenta Inception [Klie et al. 2018]. É parte também desta etapa o desenvolvimento de um plano de anotação que guiará a metodologia a ser seguida pelos anotadores.

3.3. Avaliação dos sumários

Existem métricas supervisionadas e não supervisionadas para avaliação de sumários extrativos. As métricas não supervisionadas estão mais voltadas a uma avaliação quantitativa, com foco nas separações de segmentos encontradas pelo algoritmo de sumarização. Elas serão aplicadas primeiro aos sumários obtidos.

Em posse do *gold standard dataset* gerado por especialistas, será possível tanto uma avaliação qualitativa quanto uma avaliação quantitativa considerando métricas supervisionadas. É esperado que o conjunto de sumários anotados enriqueça a avaliação e seja um diferencial no trabalho.

3.4. Execução e avaliação de classificadores sobre sumários

Os algoritmos de classificação: árvores de decisão, SVM, MLP e CNN) serão executados e avaliados sobre os sumários e sobre os textos originais. O objetivo será comparar com os resultados das abordagens de classificação que não envolvem sumários. A Figura 1 ilustra a solução a ser desenvolvida.

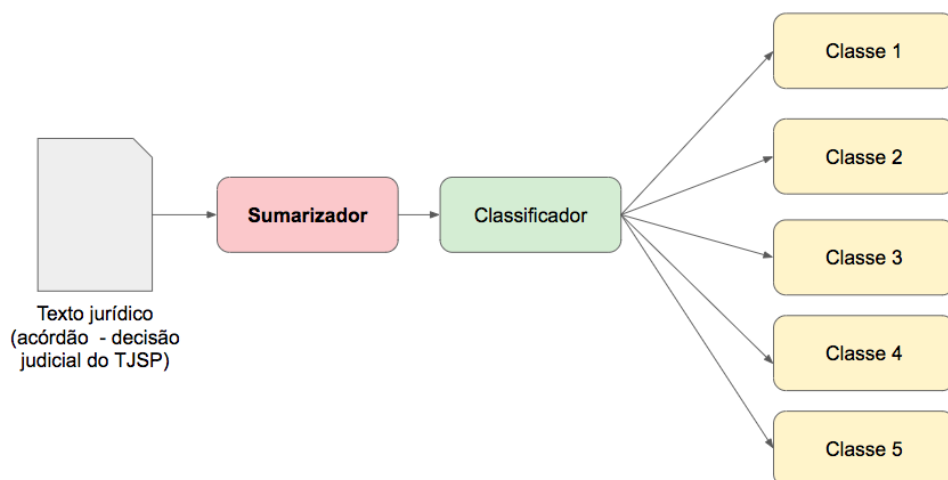


Figura 1. Etapas da solução sendo desenvolvida neste trabalho.

4. Experimentos iniciais

Foram selecionados aleatoriamente 3 acórdãos para execução dos algoritmos de sumarização selecionados. Os algoritmos LexRank, Luhn, LSA e KL-SUM foram utilizados para cada texto. Em cada execução, foi passado o parâmetro $k=5$, indicando o número de sentenças a comporem os sumários resultantes.

Observou-se uma divergência grande entre sumários gerados por diferentes algoritmos. A tabela 4 ilustra a similaridade de Jaccard obtida entre os sumários gerados pelos respectivos algoritmos, considerando os 3 acórdãos selecionados.

Tabela 1. Índice Jaccard obtido entre os sumários gerados pelos algoritmos LexRank, Luhn, KL-Sum e LSA.

	LexRank & Luhn	LexRank & KL-Sum	LexRank & LSA	Luhn & KL-Sum	Luhn & LSA	KL-Sum & LSA
Acórdão 1	0	0.25	0.25	0.1	0	0.1
Acórdão 2	0.25	0.25	0.428	0.25	0.1	0.1
Acórdão 3	0.1	0.1	0	0.1	0	0

5. Conclusão e Trabalhos Futuros

A implementação e execução dos sumários nos textos jurídicos mostrou-se efetiva e factível. Trata-se de um trabalho em andamento. Atualmente, está sendo elaborado um *corpus* anotado por especialistas para se tornar o *dataset* de referência para avaliação dos sumários.

6. Agradecimentos

Os autores agradecem o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (CNPq/MCTI/SEMPI N° 56/2022).

Referências

Chen, H., Wu, L., Chen, J., Lu, W., and Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Inf. Process. Manage.*, 59(2).

- Du, Y., Ma, T., Wu, L., Xu, F., Zhang, X., and Ji, S. (2021). Constructing contrastive samples via summarization for text classification with limited annotations. In *EMNLP*.
- El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.
- Jeong, H., Ko, Y., and Seo, J. (2016). How to improve text summarization and classification by mutual cooperation on an integrated framework. *Expert Syst. Appl.*, 60(C):222–233.
- Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).
- Rahamat Basha, S., Keziya Rani, J., and Prasad Yadav, J. J. C. (2019). A novel summarization-based approach for feature reduction enhancing text classification accuracy. *Engineering, Technology amp; Applied Science Research*, 9(6):5001–5005.
- Wang, F., Zhang, J. L., Li, Y., Deng, K., and Liu, J. S. (2021). Bayesian text classification and summarization via a class-specified topic model. *J. Mach. Learn. Res.*, 22(1).

Uso de modelagem de tópicos para agrupamento de notícias: uma abordagem usando BERTopic

Pedro Henrique Pereira¹, Ticianal Coelho da Silva¹

¹ Universidade Federal do Ceará (UFC) –Fortaleza, CE –Brasil

pedrohenripereira@gmail.com, ticianalc@insightlab.ufc.br

Abstract. *Today there is a large amount of news circulating in the media and grouping them by subjects or topics takes a lot of time. In this work, the topic modeling technique was used, through BERTopic, to group a set of news items under topics that reflect their subjects. The headlines of a set of news in English, from January to September 2022, were used. From the modeling of topics, analyzes were made and it was noticed that BERTopic can both represent the most reported topics throughout the year, as well as capture punctual events in a given period of the year. The modeling was also sensitive to variations in the context of the news.*

Resumo. *Hoje há uma grande quantidade de notícias em circulação nas mídias e agrupá-las por assuntos ou tópicos demanda muito tempo. Neste trabalho foi utilizada a técnica de modelagem de tópicos, por meio do BERTopic, para agrupar um conjunto de notícias sob tópicos que reflitam os assuntos delas. Foram utilizadas as headlines de um conjunto de notícias em inglês, de janeiro a setembro de 2022. A partir da modelagem de tópicos foram feitas análises e percebeu-se que o BERTopic tanto consegue representar os tópicos mais noticiados ao longo do ano, como também capturar eventos pontuais em um dado período do ano. A modelagem também se mostrou sensível a variações de contexto das notícias.*

1. Introdução

O advento da Web 2.0 evidenciou uma transformação nas formas de relacionamento e de interação da sociedade, influenciando também a forma como se veicula e se consome notícias. Diante da abundância de informações a que somos submetidos diariamente, muitas delas mediadas por algoritmos, torna-se essencial a identificação, a organização e a categorização dos assuntos [Arroyo-Vázquez 2014]. Esse tipo de tarefa repetitiva apresenta esforço proporcional à quantidade de dados a serem analisados e está sujeita à falha humana. Assim, surge a necessidade de automatizá-la, integral ou parcialmente, e um dos meios de fazer isso é com técnicas ligadas à área de Ciência de Dados.

Nesse contexto, uma abordagem para descobrir informações latentes em coleções de documentos é a modelagem de tópicos [Blei et al. 2003], onde um tópico representa um conjunto de palavras que descreve um assunto e os documentos são uma mistura de tópicos. Os tópicos são descobertos com base na co-ocorrência de palavras no conjunto de documentos.

Uma ferramenta que pode ser empregada nessa tarefa de modelagem de tópicos é o BERTopic, que engloba algoritmos para busca automática de tópicos representativos

em uma coleção de documentos, assumindo que documentos semanticamente semelhantes estejam em um mesmo tópico [Amorim et al. 2022].

Este trabalho busca avaliar a eficácia da ferramenta BERTopic para a tarefa de agrupar textos semanticamente afins, por meio da modelagem de tópicos. Assim, espera-se identificar, agrupados sob um mesmo tópico, notícias relacionadas por meio do assunto e também observar possíveis interações entre notícias de tópicos diferentes.

2. Fundamentação teórica

Neste trabalho foi usada a ferramenta BERTopic, desenvolvida por Maarten Grootendorst (2022), que a define como “uma ferramenta para a técnica de modelagem de tópicos que utiliza arquitetura de *transformers* e a medida c-TF-IDF para criar *clusters* densos, permitindo tópicos facilmente interpretáveis, mantendo palavras importantes nas descrições do tópico”. O BERTopic prevê 3 etapas macro: (i) *embedding* de documentos, (ii) clusterização de documentos e (iii) representação de tópicos.

3. Metodologia

O presente trabalho busca avaliar o desempenho da ferramenta de modelagem de tópicos BERTopic sobre um conjunto de textos curtos oriundos dos *headlines* de notícias de jornais. O experimento para este trabalho foi pensando em 3 etapas: (i) a coleta e preparação dos dados, (ii) a modelagem dos tópicos com o BERTopic e (iii) a análise dos tópicos.

Após o estudo do referencial teórico e de alguns trabalhos relacionados foram elaboradas três perguntas de pesquisas, a saber: (i) os tópicos gerados se alinham com as categorias atribuídas previamente pelo autor do conjunto de dados?, (ii) a variação da janela temporal dos dados influencia na formação dos tópicos? e (iii) possíveis relações no mundo real entre as notícias podem ser percebidas nos tópicos formados?

Na última etapa da metodologia proposta, na seção de Resultados, os tópicos são analisados, a fim de responder ou reestruturar tais perguntas de pesquisa.

3.1. Coleta e preparação dos dados

Para o escopo deste trabalho, foi utilizado o conjunto de *headlines* de notícias coletadas por [Misra 2022], veiculadas ao longo do ano de 2022, no site HuffPost, sendo a última publicação com data de 23/09/2022, perfazendo um total de 1.426 notícias, em inglês, que foram usadas na íntegra, sem etapa de pré-processamento. Junto às *headlines* também estava informações sobre a qual categoria cada notícia estava associada no *site*. O maior texto de *headline* tinha 103 caracteres e 18 palavras e o menor apresentava 25 caracteres e 5 palavras. A escolha por essa janela temporal foi para tornar viável a análise comparativa manual com os fatos veiculados ao longo do ano de 2022.

Para analisarmos como o tamanho da janela temporal pode influenciar na modelagem dos tópicos, optou-se por aplicar a modelagem de tópicos em dois arranjos diferentes dos dados: (i) um primeiro arranjo (A1) contendo notícias dos nove primeiros meses do ano de 2022, isto é, de janeiro a setembro e (ii) um segundo arranjo

(A2), onde as notícias serão agrupadas por trimestre: A2-trim-1 (janeiro, fevereiro e março), A2-trim-2 (abril, maio e junho) e A2-trim-3 (julho, agosto e setembro).

3.2. Modelagem dos tópicos

Por uma questão de tempo e escopo deste trabalho, foram utilizados os valores padrões sugeridos pelo BERTopic. Após a criação do modelo, a modelagem é executada recebendo as *headlines* como entrada, organizados nos conjuntos anual (A1) e trimestrais (A2-trim-1, A2-trim-2 e A2-trim-3).

4. Resultados

4.1. Primeira Análise: Alinhamento entre tópicos e categorias

Os tópicos gerados apresentam relação com as categorias atribuídas a cada notícia pelo criador do conjunto de dados. Pode-se perceber isso ao fazer uma comparação com algumas amostras da saída do experimento.

Ao verificar o conteúdo e a categoria dos documentos mais representativos do tópico 1, por exemplo, percebe-se as notícias ligadas à guerra na Ucrânia rotuladas com categorias ‘World News’ e ‘Politics’, conforme Figura 1.

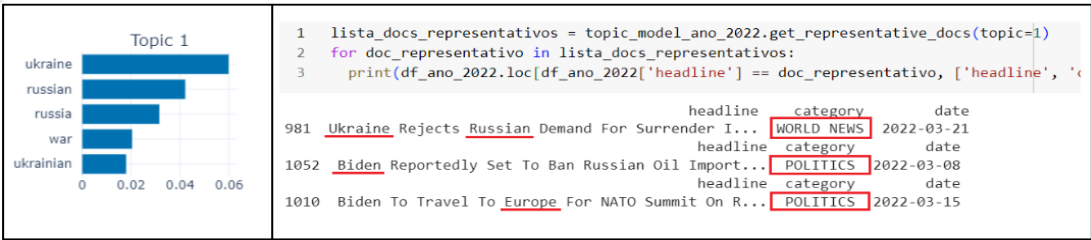


Figura 1. Conjunto de documentos mais representativos do tópico 1 da análise anual.

4.2. Segunda Análise: Comparação de tópicos em diferentes janelas temporais

Os assuntos que foram veiculados ao longo de todo o ano de 2022 geraram tópicos tanto quando se olha a modelagem anual (A1), quanto na modelagem trimestral (A2).

Isso pode ser notado nos gráficos de barra dos dois arranjos de dados, quando se observa o tópico ligado à Guerra da Ucrânia, que teve início em fevereiro de 2022 e perdura até o presente momento, junho de 2023. Na Figura 2, pode-se verificar a ocorrência desse tópico tanto quando se faz a modelagem sobre os dados de todo o ano de 2022 como quando se aplica a modelagem a cada trimestre de 2022.

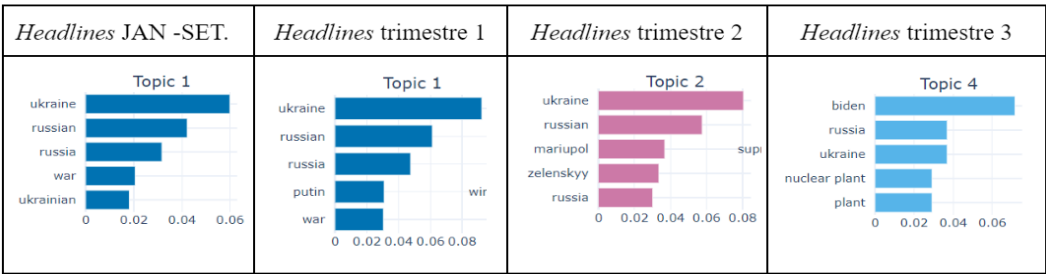


Figura 2. Tópicos ligados à Guerra da Ucrânia, nas diversas janelas temporais

Percebe-se que o assunto central do tópico permanece tanto na janela anual quanto na janela trimestral. Mas as palavras que formam o tópico vão mudando conforme se muda a janela temporal. O termo Mariupol não aparece na janela anual, mas aparece na janela correspondente ao segundo trimestre de 2022, quando houve um ataque à cidade russa de Mariupol, em 08/05/2022.

4.3. Terceira Análise: Tópicos refletem outras notícias do mundo real

A visualização de documentos proposta pelo BERTopic auxilia a perceber as relações entre tópicos e como isso é um reflexo dos acontecimentos do mundo real.

Isso é notado na visualização de documentos, no trimestre 3, ao se analisar a proximidade entre os termos ‘Biden’ e ‘Trump’, na Figura 3. Em geral, são termos que aparecem nos mesmos tipos de notícias e até mesmo juntos em muitas dessas notícias, dada a rivalidade política entre os dois. No entanto, percebe-se o termo ‘Biden’ compondo o tópico ligado à Ucrânia e distante do tópico que traz o nome do ‘Trump’.

Isso ocorreu porque, em setembro de 2022, o presidente norte-americano Joe Biden discursou na Assembleia Geral da ONU criticando fortemente o ataque russo à Ucrânia. Assim, percebe-se que para essa janela temporal apareceram mais notícias ligando o Biden à Ucrânia do que notícias relacionadas ao Trump.

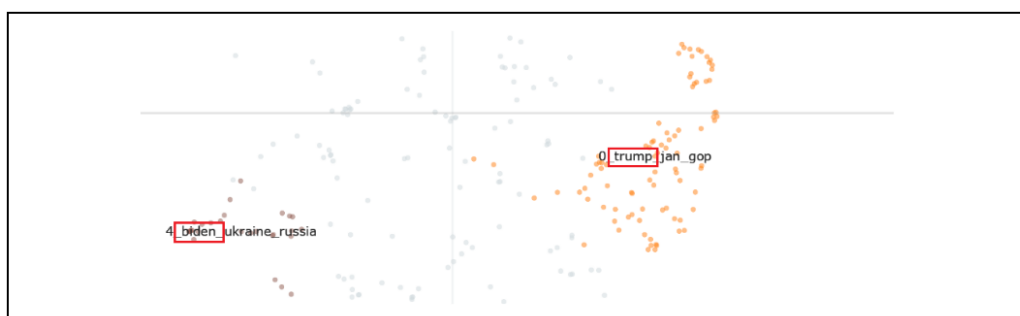


Figura 3. Visualização de documentos do trimestre 3 pelo BERTopic

5. Conclusões

Como conclusões deste trabalho, pode-se citar: (i) a metodologia proposta mostrou que o BERTopic consegue dar apoio ao processo de modelagem de tópicos, tanto na extração dos tópicos e na visualização dos resultados, de diferentes formas e (ii) os resultados mostraram que variações na janela temporal das notícias influenciaram diretamente o processo de modelagem de tópicos, uma vez que a janela temporal está ligada à quantidade de notícias de um dado assunto. Isso irá influenciar o cálculo de TF-IDF e, por conseguinte, dos tópicos gerados.

Uma contribuição é a análise na qual pode-se notar que os tópicos gerados conseguem ser mais específicos sobre os assuntos dos documentos quando comparados às categorias atribuídas pelo autor do conjunto de dados.

Um trabalho futuro seria avaliar o uso do BERTopic quando os tamanhos dos documentos apresentam grande variação de tamanhos entre si. Além do BERTopic, poderia-se também usar outras ferramentas que implementam algoritmos já difundidos em aplicações com modelagem de tópicos e *embeddings*, como, por exemplo, o Top2Vec [Angelov 2020].

Referências

- Amorim, A; Murrugarra-Llerena, N.; Silva, V.; Oliveira, D.; Paes, A. (2022). “Modelagem de Tópicos em Textos Curtos: uma Avaliação Experimental”. In: Anais do XXXVII Simpósio Brasileiro de Bancos de Dados. SBC, 2022. p. 254-266.
- Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.
- Arroyo-Vázquez, N. (2014). El content curator. Guía básica para el nuevo profesional de internet. Javier Guallar, Javier Leiva-Aguilera. Barcelona: Editorial UOC, 2013.(El profesional de la información: 24). ISBN 978-84-9064-018-0. Revista Española de Documentación Científica, v. 37, n. 2, p. e051-e051.
- Blei, D. M.; NG, A.Y.; Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, v. 3, n. Jan, p. 993-1022.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.
- Misra, R. (2022). News category dataset. arXiv preprint arXiv:2209.11429.

Proposta de avaliação da percepção dos impactos da inteligência artificial generativa na educação superior

Ana Luíza Ferreira Vieira¹, Maria Cecilia Zanon De Amorim¹, Evandro Cunha¹

¹Universidade Federal de Minas Gerais

alfv@ufmg.br, mcza@ufmg.br, cunhae@ufmg.br

Abstract. *In the early 2020s, the development and rapid dissemination of generative artificial intelligence systems, such as ChatGPT, Bing Chat and Bard, have generated debates about their potential impacts – positive and/or negative – in different areas of knowledge. In the field of higher education, these debates include discussions about possible changes in the ways of interacting with students and also of planning courses and assessments. In this work, a proposal is presented to evaluate, through the application of questionnaires, the perception of these impacts by university professors, as a way of contributing to future studies dedicated to this topic.*

Resumo. *No início da década de 2020, o desenvolvimento e a rápida disseminação de sistemas de inteligência artificial generativa, como o ChatGPT, o Bing Chat e o Bard, têm gerado debates a respeito de seus potenciais impactos – positivos e/ou negativos – em diversas áreas do conhecimento. No campo da educação superior, esses debates incluem discussões sobre possíveis mudanças nas formas de interação com os alunos e, também, de se planejar cursos e avaliações. Neste trabalho, é apresentada uma proposta de avaliação, por meio da aplicação de questionários, da percepção desses impactos por docentes do ensino superior, como forma de contribuir para futuros estudos dedicados a esse tema.*

1. Introdução

Com os recentes avanços na elaboração de modelos de linguagem em grande escala (*large language models*), sistemas de inteligência artificial generativa – como o ChatGPT (OpenAI), o Bing Chat (Microsoft) e o Bard (Google) – têm sido cada vez mais presentes na atuação de profissionais de várias áreas, inclusive no campo da educação. Nesse cenário específico, [Chen et al. 2020] mencionam que a inteligência artificial tem exercido grande impacto desde a gestão educacional até o desenvolvimento de métodos de ensino, enquanto [Tavares et al. 2020] mencionam os seguintes exemplos de aplicação de sistemas baseados em inteligência artificial: aprendizagem adaptativa, tutores inteligentes, ferramentas de diagnósticos, sistemas de recomendação, classificação de estilos de aprendizagem, mundos virtuais, gamificação e mineração de dados aplicada à educação.

Torna-se, assim, importante analisar de que forma os trabalhadores da área da educação enxergam o uso dessas tecnologias e como, segundo esses profissionais, essas ferramentas podem e/ou devem ser aplicadas em seu ambiente de trabalho. Com essa finalidade, [Parreira et al. 2021] analisam, por meio de dados obtidos via questionário

respondido por docentes universitários e do ensino médio, a percepção dos professores sobre as inovações tecnológicas em seu campo de atuação. Dentre os resultados obtidos, destaca-se a atitude positiva, de maneira geral, dos docentes perante as inovações tecnológicas "de primeira geração", além de uma consciência de que sistemas de inteligência artificial exercerão um impacto importante no futuro. Verificou-se, ainda, que a percepção desses impactos varia conforme fatores demográficos como idade e nível de escolaridade.

Além do mais, vale destacar a importância do desenvolvimento de habilidades e competências por parte dos profissionais da educação para lidar com as mais diversas ferramentas de inteligência artificial que surgem a cada momento, a fim de extraírem destas, contribuições positivas e seguras para o processo de ensino-aprendizagem. Desse modo, se faz necessário que mais trabalhos acerca do tema contribuam para tal entendimento.

Entretanto, a percepção dos impactos dos sistemas de inteligência artificial generativa (como o ChatGPT, o Bing Chat e o Bard) na educação ainda não foram suficientemente estudados, em especial no Brasil. Este trabalho busca contribuir para esse debate ao propor um instrumento específico para essa finalidade. O objetivo é permitir a análise do ponto de vista dos professores universitários em relação ao uso de sistemas de inteligência artificial generativa como ferramentas de apoio ao ensino e aprendizagem. É proposto um questionário que possa oferecer subsídios para a compreensão de como esses profissionais percebem e utilizam essas tecnologias para o aprendizado autônomo dos estudantes, além de compreender como elas afetam o usuário e o processo de aprendizagem, assim como definir o posicionamento dos docentes sobre o uso dessa ferramenta.

2. Metodologia e proposta de questionário

A proposta aqui apresentada é um questionário a ser administrado a professores do ensino superior brasileiro. Esse questionário foi elaborado a partir de perguntas e hipóteses de pesquisa surgidas, sobretudo, de três situações: (a) interações com docentes; (b) levantamento bibliográfico; (c) discussões internas no grupo de pesquisa. Em particular, foi levantada a hipótese de existir certa hostilidade em relação ao uso dessas ferramentas por parte do público estudado, a qual poderia ser mais ou menos acentuada em função da área de atuação (por exemplo, entre docentes de áreas tradicionalmente mais ou menos voltadas para o uso de novas ferramentas tecnológicas), da idade e das habilidades com computadores do docente.

Outras perguntas que guiaram a elaboração do questionário foram: os docentes acreditam que o uso de sistemas de inteligência artificial generativa possa auxiliar na formação de seus alunos (talvez principalmente em algumas áreas específicas)? Há a percepção de que o uso dessas ferramentas aumente a autonomia do discente? Há, por outro lado, a percepção de que esses sistemas possam ser utilizados como uma forma de burlar e ludibriar o professor? Existe receio de que essas ferramentas possam, eventualmente, substituir os professores na universidade? Para responder a essas e a outras questões relacionadas, a etapa seguinte deste projeto de pesquisa envolve enviar o questionário proposto a docentes da educação superior que atuam em instituições de ensino superior brasileiras, em diversas áreas do conhecimento.

O questionário apresentado neste trabalho possui aspecto qualiquantitativo e inclui uma seção para obtenção de informações demográficas (idade, gênero, escolaridade, área de atuação), questões referentes ao contato com ferramentas computacionais

e questões diretamente voltadas para a indicação da percepção do docente sobre os impactos da inteligência artificial generativa na educação superior. Nesta última seção, algumas questões serão aferidas por meio de escala de Likert, de 1 (discordo totalmente) a 5 (concordo totalmente) [Dalmoro. M 2013], enquanto outras são perguntas de resposta aberta. Além disso, para a criação do questionário, foi levado em consideração a proposta de [Faleiros et al. 2016] que apresenta as limitações de um questionário online, como a exclusão de analfabetos digitais e o impedimento do auxílio ao participante quando o mesmo não compreende alguma pergunta. Apesar dessas limitações, decidiu-se pelo uso do questionário online por favorecer o alcance de docentes de outras universidades, de outros idiomas e com uma velocidade maior. Assim, a proposta de questionário é apresentada a seguir.

1. Qual é a sua idade?
(a) Menos de 30 anos (b) De 30 a 39 anos (c) De 40 a 49 anos
(d) De 50 a 59 anos (e) De 60 a 69 anos (f) A partir de 70 anos
 2. Com qual gênero você se identifica?
(a) Feminino (b) Masculino (c) Prefiro não informar (d) Outro: _____
 3. Qual é o seu grau mais alto de escolaridade (completo)?
(a) Ensino superior (b) Especialização (c) Mestrado (d) Doutorado
 4. Qual é a sua área de atuação na educação superior? *Por exemplo: Linguística, Literatura, Ciência da Computação, Medicina etc. Informe quantas áreas desejar.*
 5. De maneira geral, como você define o seu grau de habilidade/intimidade com computadores e ferramentas tecnológicas? *As respostas a esta questão serão aferidas por meio de escala de Likert, de 1 (muito baixo) a 5 (muito alto).*
 6. Você conhece os sistemas de inteligência artificial generativa (como o ChatGPT, o Bing Chat ou o Bard)?
(a) Não conheço (b) Já ouvi falar, mas nunca usei (c) Conheço e já usei por curiosidade e/ou para finalidades não acadêmicas (d) Conheço e já usei para finalidades acadêmicas
 7. Caso já tenha utilizado algum desses sistemas, qual grau de dificuldade você encontrou durante o processo? *Considere a dificuldade para acessar o sistema, para interagir com ele, para fazê-lo reagir da forma desejada etc.*
(a) Dificuldade baixa (b) Dificuldade média (c) Dificuldade alta
 8. Caso já tenha utilizado algum desses sistemas, quais recursos você explorou? *Marque quantas opções desejar.* (a) Pesquisa sobre um tópico específico (b) Geração de texto sobre um tópico específico (c) Obtenção de resposta a uma dúvida (d) Correção/revisão textual (e) Sumarização/resumo (f) Tradução (g) Diálogo (h) Outro: _____
- As respostas às questões 9 a 20 serão aferidas por meio de escala de Likert, de 1 (discordo totalmente) a 5 (concordo totalmente).**
9. Na média, sistemas de inteligência artificial generativa, como o ChatGPT e similares, são ferramentas de fácil acesso e uso para meus estudantes atuais.
 10. Eu me sinto à vontade/confortável com a possibilidade do uso crescente de sistemas de inteligência artificial generativa pelos meus estudantes.
 11. Considero positivo que meus estudantes utilizem sistemas de inteligência artificial generativa para *estudar conteúdos* referentes às disciplinas que leciono.
 12. Considero positivo que meus estudantes utilizem sistemas de inteligência artificial generativa para *redigir textos* referentes às disciplinas que leciono.

13. De maneira geral, acredito que sistemas de inteligência artificial generativa podem facilitar/auxiliar no aprendizado dos estudantes na minha área de atuação.
14. Eu me sinto preparado/a para responder aos novos desafios que sistemas de inteligência artificial generativa podem ocasionar na educação superior, na minha área de atuação.
15. **A curto prazo (em menos de três anos)**, acredito que esses sistemas serão capazes de substituir, parcialmente ou integralmente, professores da minha área de atuação na educação superior.
16. A curto prazo (em menos de três anos), acredito que esses sistemas serão capazes de realizar, com bom aproveitamento, as atividades (provas, trabalhos etc.) que, atualmente, eu solicito aos meus estudantes.
17. A curto prazo (em menos de três anos), acredito que será necessário replanejar as disciplinas que leciono (plano de ensino, atividades avaliativas etc.) em função dos sistemas de inteligência artificial generativa.
18. **A longo prazo (em mais de dez anos)**, acredito que esses sistemas serão capazes de substituir, parcialmente ou integralmente, professores da minha área de atuação na educação superior.
19. A longo prazo (em mais de dez anos), acredito que esses sistemas serão capazes de realizar, com bom aproveitamento, as atividades (provas, trabalhos etc.) que, atualmente, eu solicito aos meus estudantes.
20. A longo prazo (em mais de dez anos), acredito que será necessário replanejar as disciplinas que leciono (plano de ensino, atividades avaliativas etc.) em função dos sistemas de inteligência artificial generativa.

As questões 21 e 22 são perguntas de resposta aberta.

21. Quais são suas expectativas, a curto e/ou longo prazo, com relação ao uso de sistemas de inteligência artificial generativa (como o ChatGPT, o Bing Chat e o Bard, por exemplo) no contexto da educação superior brasileira, em particular na sua área de atuação?
22. Você gostaria de adicionar algum comentário ou opinião referente ao tema desta pesquisa?

3. Conclusão e trabalhos futuros

A elaboração do questionário qualiquantitativo apresentado neste trabalho tem como objetivo fornecer um meio padronizado de se avaliar a percepção dos impactos da inteligência artificial generativa na educação superior. A finalidade do projeto de pesquisa em que este trabalho se insere é investigar essa percepção entre docentes universitários no Brasil – por essa razão, o questionário proposto será administrado a esses sujeitos. No entanto, pode ser interessante administrá-lo também a docentes universitários estabelecidos em outros países, para fins de comparação dos resultados obtidos.

Ao compreender melhor como os docentes percebem e utilizam os sistemas de inteligência artificial generativa, como o ChatGPT, o Bing Chat e o Bard, será possível obter informações relevantes para o desenvolvimento, a implementação e o aprimoramento dessas ferramentas no contexto educacional. Após a aplicação dos questionários, espera-se que a análise dos dados coletados permita identificar padrões, tendências e desafios relacionados ao uso dessas tecnologias na educação, fornecendo uma base para futuras pesquisas na área.

Entre os trabalhos futuros previstos no projeto de pesquisa no qual este trabalho se insere, se destacam o desenvolvimento de outros dois questionários, destinados a estu-

dantes do ensino superior e professores da educação básica. Em ambos os casos, a finalidade será compreender até que ponto as percepções desses públicos sobre os impactos da inteligência artificial generativa na educação são similares ou divergentes daquelas dos docentes do ensino superior.

References

- Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: a review. *IEEE Access*, 8:75264–75278.
- Dalmo, M. V. K. (2013). Dilemas na construção de escalas tipo likert: o número de itens e a disposição influenciam nos resultados? *REVISTA GESTÃO ORGANIZACIONAL — VOL. 6 - EDIÇÃO ESPECIAL*.
- Faleiros, F., K  ppler, C., Pontes, F. A. R., Silva, S. S. d. C., Goes, F. d. S. N. d., and Cuckick, C. D. (2016). Use of virtual questionnaire and dissemination as a data collection strategy in scientific studies. *Texto & Contexto - Enfermagem*, 25(4):e3880014.
- Parreira, A., Lehmann, L., and Oliveira, M. (2021). O desafio das tecnologias de intelig  ncia artificial na Educa  o: percep  o e avalia  o dos professores. *Ensaio: Avalia  o e Pol  ticas P  blicas em Educa  o*, 29(113):975–999.
- Tavares, L. A., Meira, M. C., and Amaral, S. F. d. (2020). Intelig  ncia artificial na educa  o: survey. *Brazilian Journal of Development*, 6(7):48699–48714.

Explorando Redes Neurais Profundas para Tarefa de Aceitabilidade Linguística

Henrique Santos¹, Késia P. Alencar², Rogério F. de Sousa², Rafael T. Anchiêta²

¹Universidade Federal do Piauí – Picos (UFPI)

²Instituto Federal do Piauí – Picos (IFPI)

Laboratório de Inteligência Artificial, Robótica e Automação – (LIARA)

{henriques.santod, kesiapereira2003}@gmail.com

{rogerio.sousa, rta}@ifpi.edu.br

Abstract. *Linguistic acceptability is the task of determining whether a sentence is grammatically correct. Although some grammar correction tools exist for Portuguese, they are based on manually defined rules, which is a laborious task. This work investigated deep neural networks and language models for the linguistic acceptability task to develop more robust tools/methods for Portuguese to obtain better results than the existing tools. Recurrent networks, convolutional networks, and the BERTimbau and Albertina language models were explored. These models were trained on a corpus translated from English to Portuguese and evaluated on the Probi corpus. The recurrent and convolutional networks achieved the best results (0.37 f1), being competitive with the LanguageTool tool (0.40 f1).*

Resumo. *Aceitabilidade linguística é a tarefa de determinar se uma sentença está gramaticalmente correta. Apesar de existir algumas ferramentas de correção gramatical para o português, elas são baseadas em regras manualmente definidas, o que é uma tarefa laboriosa. Neste trabalho, investigaram-se redes neurais profundas e modelos língua para a tarefa de aceitabilidade linguística a fim de desenvolver ferramentas/métodos mais robustos para o português que obtenham resultados melhores do que as ferramentas existentes. Explorou-se redes recorrentes, redes convolucionais e os modelos de língua BERTimbau e Albertina. Esses modelos foram treinados em um corpus traduzido do inglês para o português e avaliados no corpus Probi. As redes recorrentes e convolucionais atingiram os melhores resultados (0,37 f1), sendo competitivas com a ferramenta LanguageTool (0,40 f1).*

1. Introdução

Aceitabilidade linguística é a tarefa de determinar se uma sentença está gramaticalmente correta. Essa tarefa vem do campo da linguística generativa [Klezi et al. 2022] que se baseia em julgamentos intuitivos de falantes nativos sobre se uma sentença é aceitável ou não [T Schütze 2016]. Essa tarefa possui diversas aplicações na área de Processamento de Línguas Naturais (PLN), por exemplo: analisar a robustez de modelos de língua [Yin et al. 2020] e verificar se tais modelos adquirem conhecimentos gramaticais [Zhang et al. 2021, Choshen et al. 2022]. Além disso, aceitabilidade linguística tem

sido utilizada para avaliar a correção gramatical e a semântica de métodos de geração automática de textos [Bakshi et al. 2021, Batra et al. 2021].

Embora aceitabilidade linguística seja uma área de estudo com aplicações recentes focadas, principalmente, em modelos de língua, ela é pouco explorada na língua portuguesa. Acredita-se que seja devido à falta de um *corpus* manualmente anotado. Apesar da falta desse recurso, existem algumas ferramentas que fazem correção ortográfica e gramatical na língua portuguesa, por exemplo: CoGrOO [Kinoshita et al. 2007] e LanguageTool¹. Essas ferramentas, fazem as correções em um texto através de regras linguísticas manualmente projetadas. Métodos baseados em regras, embora apresentem bons resultados, são dependentes de especialistas para a construção de regras. Ademais, a criação de regras denota muito esforço.

Com o objetivo de diminuir o esforço na criação de regras linguísticas, neste trabalho exploraram-se métodos de aprendizado profundo (*deep learning*) e modelos de língua para a identificação automática de sentenças aceitáveis. Esses métodos mapeiam uma entrada para a saída desejada de maneira automática através de uma grande quantidade de dados sem a necessidade de desenvolver regras manuais.

A fim de treinar os modelos para a tarefa de aceitabilidade linguística, utilizou-se a versão traduzida do *corpus* CoLA (*Corpus of Linguistic Acceptability*) [Warstadt et al. 2019] para o português. O *corpus* CoLA é um conjunto de 1.657 sentenças em inglês extraídas da literatura linguística e rotuladas como gramatical ou agramatical. A versão portuguesa² do *corpus* CoLA foi traduzida automaticamente utilizando ferramenta *Google Cloud Translation*. Após o treinamento dos modelos, eles foram avaliados no *corpus* Probi [Martins 2002] que é um recurso composto por 11.625 sentenças anotadas manualmente. Do total de sentenças, 2.616 são anotadas como agramatical, ou seja, possuem algum tipo de erro. Obteve-se o melhor resultado com uma rede convolucional, atingido 0,37 de medida F1 que é um valor competitivo com a ferramenta LanguageTool que atingiu 0,40 na mesma métrica.

O restante do artigo está organizado da seguinte forma. Seção 2 introduz brevemente os trabalhos relacionados. Na Seção 3, é detalhado os métodos desenvolvidos. Seção 4 apresenta os resultados e discussões. Por fim, Seção 5 conclui o artigo indicando futuras direções.

2. Trabalhos Relacionados

O ReGra foi o primeiro revisor gramatical para a língua portuguesa [Nunes and Jr. 2000]. Ele foi desenvolvido através de regras manualmente definidas. Posteriormente, o revisor foi incorporado ao MS Word e ao processador de texto REDADOR da Itaotec.

CoGrOO (Corretor Gramatical para OpenOffice) [Kinoshita et al. 2007] é uma ferramenta de correção gramatical para a língua portuguesa. Assim como o ReGra, o corretor gramatical foi desenvolvido através de regras gramaticais manualmente definidas.

LanguageTool é ferramenta de correção gramatical disponível para vários idiomas, como: inglês, alemão, espanhol, holandês, português, entre outros. Da mesma forma

¹<https://languagetool.org/>

²<https://github.com/ju-resplande/PLUE>

que as ferramentas anteriores, o módulo de correção gramatical para língua portuguesa é baseado em regras.

Percebe-se que as principais ferramentas de correção gramatical para o português são baseados em regras linguísticas. Nossa abordagem investiga métodos baseados em redes neurais profundas e modelos de língua como uma alternativa aos métodos tradicionais baseados em regras. No que segue, detalha-se os métodos desenvolvidos.

3. Métodos Desenvolvidos

A fim de desenvolver um método de aceitabilidade linguística para a língua portuguesa, investigaram-se as Redes Neurais Convolucionais (CNN), Redes Neurais Recorrentes (RNN) e os modelos de língua BERTimbau [Souza et al. 2020] e Albertina [Rodrigues et al. 2023]. Esses modelos foram treinados ou ajustados no *corpus* CoLA traduzido para a língua portuguesa e testados no *corpus* Probi que foi criado para avaliar o revisor gramatical ReGra. Analisou-se diversas configurações para os modelos supracitados. Os melhores parâmetros para a RNN e CNN são apresentados nas Tabelas 1 e 2, respectivamente.

Em ambas as redes neurais, utilizou-se as embeddings pré-treinadas do GloVe de 300 dimensões [Hartmann et al. 2017]. Para a RNN, adotou-se uma rede *Bidirectional Long Short-Term Memory* (BiLSTM). Para a CNN, utilizou-se três camadas de convolução e cem filtros com dimensões de três, quatro e cinco. As final da RNN e CNN, implementou-se uma camada linear para realizar a classificação.

Tabela 1. Parâmetros rede RNN.

Parâmetro	Valor
Embeddings	300
Dropout	0,5
Bidirecional	sim
Otimizador	Adam
Erro	BCE
Épocas	2

Tabela 2. Parâmetros rede CNN.

Parâmetro	Valor
Embeddings	300
Dropout	0,1
Convolução	3 camadas
Filtros#	100
Tamanho dos filtros	[3, 4, 5]
Otimizador	Adam
Erro	BCE
Épocas	2

Para o modelo de língua Albertina, investigou-se algumas variações de parâmetros. No entanto, os resultados sempre foram iguais. Na Tabela 3, são apresentados os parâmetros desse modelo. Diferente do Albertina, os resultados do BERTimbau variaram de acordo com o parâmetro definido. Na Tabela 4, são apresentados os parâmetros que apresentaram melhores resultados.

O código fonte para o treinamento e ajuste fino dos modelos está disponível em <https://github.com/liara-ifpi/aceitabilidade>. Na seção seguinte, os resultados dos modelos desenvolvidos são apresentados.

4. Resultados e Discussão

A partir dos modelos implementados, realizou-se uma comparação entre eles. Além disso, também comparou-se os resultados com as ferramentas CoGrOO e LanguageTool. Na

Tabela 3. Parâmetros Albertina.

Parâmetro	Valor
Dropout	0,5
Batch size	8
Otimizador	AdamW
Aprendizagem	4×10^{-5}
Erro	CrosEntropy
Épocas	6

Tabela 4. Parâmetros BERTimbau.

Parâmetro	Valor
Batch size	8
Otimizador	AdamW
Aprendizagem	4×10^{-5}
Erro	CrossEntropy
Épocas	2

Tabela 5, são apresentados os resultados obtidos.

Tabela 5. Melhores resultados alcançados.

Modelo/Ferramenta	Precisão	Cobertura	F1
CoGrOO	0,59	0,19	0,28
LanguageTool	0,56	0,32	0,40
Albertina	0,00	0,00	0,00
BERTimbau	0,36	0,16	0,22
RNN	0,24	0,78	0,37
CNN	0,23	1,00	0,37

A partir dos resultados, pode-se ver que o modelo Albertina teve o pior resultado. O Albertina é um modelo de língua para o português recente, que nas avaliações realizadas por seus desenvolvedores atingiu resultados melhores que o BERTimbau. Em trabalhos futuros será investigado porque nesta tarefa os resultados foram muito inferiores ao BERTimbau. O CoGrOO e LanguageTool atingiram os melhores resultados em precisão e f1, respectivamente. Nossos melhores modelos foram a RNN e a CNN, ambas alcançaram 0,37 na medida f1. Além disso, a CNN atingiu 1,00 na cobertura.

Os resultados alcançados tanto pela RNN quanto pela CNN são, de certa forma, promissores, pois os modelos foram treinados em um *corpus* que foi traduzido de forma automática para o português. Isso pode justificar a baixa precisão dos modelos, ou seja, uma alta taxa de falsos positivos. Um trabalho futuro é investigar a qualidade da tradução automática nesse *corpus*.

5. Conclusão

Neste trabalho, explorou-se redes neurais profundas e modelos de língua para a tarefa de aceitabilidade linguística. Apesar de existir algumas ferramentas de correção gramatical para o português, elas são baseadas de regras manualmente definidas. Este artigo buscou uma forma alternativa para identificar se uma sentença é aceitável, ou seja, se está gramaticalmente correta. Os melhores resultados obtidos foram com uma rede BiLSTM CNN, atingindo um resultado competitivo com a ferramenta LanguageTool e superando a ferramenta CoGrOO.

Como trabalho futuro, além de explorar mais o modelo de língua Albertina que obteve os piores resultados e investigar o *corpus* CoLA traduzido para o português, pretende-

se realizar uma detalhada análise de erros a fim de descobrir onde os modelos erraram, desenvolver um método que possa capturar os pontos fortes do BERTimbau e da CNN para melhorar os resultados e uma ferramenta que online para uso dos modelos desenvolvidos.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq pelo apoio financeiro.

Referências

- Bakshi, S., Batra, S., Heidari, P., Arun, A., Jain, S., and White, M. (2021). Structure-to-text generation with self-training, acceptability classifiers and context-conditioning for the GEM shared task. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 136–147, Online. Association for Computational Linguistics.
- Batra, S., Jain, S., Heidari, P., Arun, A., Youngs, C., Li, X., Donmez, P., Mei, S., Kuo, S., Bhardwaj, V., Kumar, A., and White, M. (2021). Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Choshen, L., Hacohen, G., Weinshall, D., and Abend, O. (2022). The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.
- Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.
- Kinoshita, J., Salvador, L., Menezes, C., and Silva, W. (2007). Cogroo - an openoffice grammar checker. In *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, pages 525–530, Rio de Janeiro, Brazil. IEEE.
- Klezl, J., Mohammed, Y. A., and Volodina, E. (2022). Exploring linguistic acceptability in Swedish learners’ language. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 84–94, Louvain-la-Neuve, Belgium. LiU Electronic Press.
- Martins, R. (2002). Probi: um corpus de teste para o revisor gramatical regra. Technical report, ICMC-USP.
- Nunes, M. d. G. V. and Jr., O. (2000). O processo de desenvolvimento do revisor gramatical regra. In *XXVII Seminário Integrado de Software e Hardware*, pages 1–15, Curitiba, 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-*.

- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *9th Brazilian Conference on Intelligent Systems*, pages 403–417, Rio Grande, Brazil. Springer.
- T Schütze, C. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.
- Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Yin, F., Long, Q., Meng, T., and Chang, K.-W. (2020). On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3386–3403, Online. Association for Computational Linguistics.
- Zhang, Y., Warstadt, A., Li, X., and Bowman, S. R. (2021). When do you need billions of words of pretraining data? In *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 1112–1125, Online. Association for Computational Linguistics.