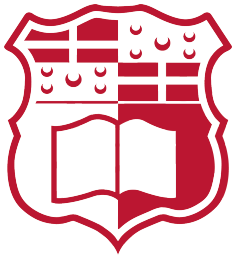


Lessons learned from the evaluation of Portuguese Language Models

Ruan Chaves Rodrigues

Supervised by Dr Marc Tanti

Co-supervised by Dr Rodrigo Agerri

Department of Artificial Intelligence

Faculty of Information & Communication Technology

University of Malta

September, 2023

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. in Human Language Science and Technology.

www.um.edu.mt

First edition, September 28, 2023

To the ocean

my constant companion as these pages came to life

Acknowledgements

I extend my deep appreciation to the Portuguese natural language processing community, whose collective efforts laid the foundation for this study. I'd like to especially acknowledge Francielle Vargas, Hugo Gonçalo Oliveira, and Juliana Resplande Sant'Anna Gomes for their assistance and insights regarding Portuguese benchmarks and datasets.

I am also indebted to the academic community at the University of Malta. Their rigorous feedback significantly enriched the depth of my work. Paul Marty's insights into our statistical analysis were indispensable, and I am grateful to Marc Tanti for his guidance as my supervisor at UM.

Special gratitude is reserved for Rodrigo Agerri, my supervisor at the University of the Basque Country. His research paper, "Lessons Learned from the Evaluation of Spanish Language Models", co-authored with Eneko Agirre, served as the primary inspiration for my research. Our orientation meetings were fundamental in defining the current shape of this study.

I am grateful to the Ixa group at UPV/EHU and the German Research Center for Artificial Intelligence (DFKI). Their support in providing access to their GPU clusters was crucial for our experiments. I am also immensely grateful to the people and institutions involved in the European Masters Program in Language & Communication Technologies (LCT). They generously supported my research journey with an Erasmus scholarship and assisted me with the intricacies of living abroad as a Master's student.

Finally, a personal thank you goes to my grandmother, Magna, who supported me at every step of my education.

Abstract

With the rising prominence of the Portuguese language in Natural Language Processing (NLP), a clear divide is observed between major corporations and smaller academic entities in model training. This raises an important question: can the efforts of smaller entities compete with major corporations in Portuguese natural language tasks? And which aspects should they prioritize to enhance their advantage? In our pursuit to answer this, we provide a historical overview of advancements in Portuguese NLP, from early word embeddings to the rise of Large Language Models (LLMs). We then discuss the linguistic challenges of benchmark construction and set out to perform a comprehensive evaluation of modern language models using a carefully designed benchmark. Using detailed evaluation methods and rigorous statistical analysis, our findings show no significant performance differences between models trained solely on Portuguese datasets and those trained on multilingual data. Our study challenges the perceived benefits of current Portuguese language models and highlights the need for deeper linguistic research and evaluation in Portuguese NLP. Our main contribution, the Natural Portuguese Language Benchmark (Napolab), is available at `https://github.com/ruanchaves/napolab`.

Contents

List of Figures

List of Tables

List of Abbreviations

ABSAPT 2022 Aspect-Based Sentiment Analysis in Portuguese, 2022 Edition
AIA-BDE Apoio Inteligente a Empreendedores - Balcão do Empreendedor (Intelligent Support for Entrepreneurs - Entrepreneur's Desk)
API Application Programming Interface
ASO Almost Stochastic Order
ASSIN Avaliação de Similaridade Semântica e INferência textual (Semantic Similarity and Textual Inference Evaluation)
BERT Bidirectional Encoder Representations from Transformers
BLEU Bilingual Evaluation Understudy
BrWaC Brazilian Portuguese Web as Corpus
CC-News Common Crawl News
CC100 Common Crawl 100
DIP Desafio de identificação de personagens (Character Identification Challenge)
F_{train} Word Train Frequency
GLUE General Language Understanding Evaluation
GPU Graphics Processing Unit
HAREM HAREM é uma Avaliação de Reconhecedores de Entidades Mencionadas (HAREM is an Evaluation of Recognizers of Mentioned Entities)
IDPT 2021 Irony Detection in Portuguese, 2021 Edition
ID Identifier
IXAes IXABERTesv2
IberLEF Iberian Languages Evaluation Forum
LC Label Consistency
len Length
LLM Large Language Model
LeNER-BR Legal Named Entity Recognition for Brazilian Portuguese
METEOR Metric for Evaluation of Translation with Explicit ORdering
NER Named Entity Recognition
NLI Natural Language Inference
NLLB No Language Left Behind

NLM Neural Language Model
NLP Natural Language Processing
Napolab Natural Portuguese Language Benchmark
OOV Out-of-Vocabulary
OSCAR Open Super-large Crawled Aggregated coRpus
PLM Pretrained Language Model
PLUE Portuguese Language Understanding Evaluation
pss PorSimplesSent (Portuguese Simple Sentences)
PT-BR Portuguese from Brazil
PT-PT Portuguese from Portugal
PTBR Portuguese from Brazil
PTPT Portuguese from Portugal
Poeta Portuguese Evaluation Tasks
PorSimplesSent Portuguese Simple Sentences
QNLI Question-Answering Natural Language Inference
R_{OOV} Rate of Out-of-Vocabulary Tokens
RTE Recognizing Textual Entailment
R_{WO} Rate of Word Overlap
ReLi Resenha de Livros (Book Reviews)
ReRelEM Reconhecimento de Relações entre Entidades Mencionadas (Recognition of Relations between Mentioned Entities)
SA Sentiment Analysis
SICK Sentences Involving Compositional Knowledge
SIGARRA Sistema de Informação para Gestão Agregada dos Recursos e dos Registos Académicos (Information System for Aggregated Management of Academic Resources and Records)
SOTA State-of-the-Art
SQuAD Stanford Question Answering Dataset
STM Statistical Language Model
STS Semantic Textual Similarity
XML eXtensible Markup Language

1

Introduction

Given the widespread importance of the Portuguese language, numerous groups have embraced the challenge of training and evaluating NLP models for Portuguese tasks. The players in this field can be broadly divided into two groups. On one hand, there are large companies that possess the resources to train massive state-of-the-art models and often develop multilingual models capable of handling Portuguese tasks alongside other languages. For instance, Google has actively contributed to this field with models like the BERT Multilingual model Devlin (2020), Meta has developed the XLM-RoBERTa model Conneau et al. (2020), and Microsoft has developed the DeBERTa models He et al. (2020, 2023), all of them multilingual models capable of performing Portuguese NLP tasks.

On the other hand, there are groups with fewer resources that tend to focus on developing models exclusively for the Portuguese language. These groups can be broadly divided into academic groups, startups, and open-source community initiatives. Notable academic groups include the NILC, a research group based in Brazil, which, for instance, released Portuguese static word embeddings Hartmann et al. (2017); Deep Learning Brazil, which released Portuguese ELMo models de Castro et al. (2019); and NLX - Natural Language and Speech Group, based in Portugal, which released a Portuguese DeBERTa model Rodrigues et al. (2023). There are also startups such as Neuralmind, which released Portuguese T5 and BERT models Carmo et al. (2020); Souza et al. (2020), and Maritaca.ai, which trained Portuguese LLMs based on GPT-J and LLaMA Pires et al. (2023). Finally, there are open-source non-profit initiatives such as the 22h Group, which pre-trained an LLaMA model on Portuguese data Larcher et al. (2023).

Turning our attention to the efforts of large companies, the issue of the "curse of multilinguality" has been extensively discussed in relation to modern Transformer-based multilingual models Conneau et al. (2019). When a model supports more than a handful of languages, its performance on each language tends to suffer. This has led to two distinct lines of research: either improving the model architecture itself to mitigate the curse of multilinguality while maintaining multi-language support

Pfeiffer et al. (2022); or giving up on the multilingual approach and focusing on training better models for just one target language using existing architectures.

The second avenue, focusing on a specific target language, has been the preferred approach for academic groups and smaller companies working on Natural Language Processing for Portuguese. However, it comes with its own set of challenges. Ducel et al. (2022) have observed that NLP research centered around the English language tends to receive greater acceptance at prestigious publication venues compared to research focused on other languages. This is because such work might be perceived as "language-specific" and thus considered less pertinent. This skewed incentive structure for researchers contributes to a situation where languages other than English are more likely to be neglected in the development of new models.

Given the restricted allocation of resources—both in terms of human effort and computational power—for the development of models for specific languages, it's possible that models pretrained for a particular language might not attain the same level of performance as large multilingual models pretrained on a multitude of languages, despite of the curse of multilinguality. This situation has been evident, for instance, in the case of the Spanish language: Agerri and Agirre (2022) have demonstrated that the multilingual XLM-RoBERTa model, previously overlooked in the literature, outperformed monolingual models pretrained exclusively on Spanish data in downstream tasks. This was achieved through a comprehensive evaluation of the performance of various models on a diverse set of tasks on Spanish datasets.

In this study, significantly motivated by the lack of analogous investigations for the Portuguese language, our objective is to conduct a thorough evaluation of Portuguese language models and multilingual models. We will also assess the performance of these models across a diverse array of tasks using meticulously curated Portuguese datasets, which include tasks that have not been previously explored in the literature. In doing so, we aspire to provide researchers in Natural Language Processing for the Portuguese language with valuable insights regarding the training strategies most likely to yield optimal results.

1.1 | Motivation

While a plethora of Transformer models pretrained exclusively on general-purpose Portuguese data can be readily found online, it's worth noting that only two of these models have undergone thorough evaluation in scientific literature and have been pretrained at a scale comparable to their multilingual counterparts. These two notable models are the BERTimbau models, introduced by Souza et al. (2020), and the more recent Albertina models, presented by Rodrigues et al. (2023).

Despite the widespread mentions and applications of the BERTimbau models across various contexts within the literature, minimal efforts have been dedicated to assessing their performance in comparison to multilingual models. Moreover, the existing Portuguese models have not undergone evaluations with the same level of

rigor as their multilingual counterparts introduced thus far. A critical concern arises from the fact that a significant portion of evaluation endeavors focus on datasets automatically translated from English to Portuguese. While this resolves the issue of data scarcity, it simultaneously introduces a new challenge: these translated datasets tend to exhibit a distinct language register referred to as "translationese" Chowdhury et al. (2022); Edunov et al. (2019), which may deviate considerably from the Portuguese language as it naturally occurs in real-world tasks.

Hence, fostering the development of competitive, all-encompassing general-purpose models tailored for tackling Natural Language Processing tasks in Portuguese requires a comprehensive evaluation of the existing models. This evaluation should encompass a comparative analysis against multilingual alternatives, spanning a diverse array of tasks. It should also rely on meticulously curated Portuguese datasets that closely simulate the challenges encountered in actual production environments. This study stands as an indispensable endeavor motivated by these considerations.

1.2 | Aims and Objectives

This research aims to provide a comprehensive evaluation of Portuguese language models and multilingual models, allowing for a reliable assessment of their performance. To achieve this goal, we have defined three specific objectives that pertain to the evaluation of these models, the utilization of automatic translation, and the selection of appropriate datasets.

1.2.1 | First Objective

> *Constructing a comprehensive evaluation benchmark for Portuguese language models and multilingual models that can consistently assess the performance of each individual model.*

This objective is pursued through a meticulous process of dataset selection, guided by the principles established in prior benchmarks such as GLUE and SuperGLUE Wang et al. (2018, 2019). Our approach ensures not only the reproducibility of results but also their alignment with real-world performance scenarios of the models.

1.2.2 | Second Objective

> *Thoroughly evaluating the performance of Portuguese language models and multilingual models across a diverse array of tasks.*

This objective is accomplished by devising a model fine-tuning and evaluation pipeline rooted in recent research on optimal practices for fine-tuning Transformer models and deep learning architectures in general. The pipeline encompasses three

stages: hyperparameter optimization, the identification of optimal random seeds, and the subsequent fine-tuning of models for a specified number of epochs.

1.2.3 | Third Objective

> *Assessing the impact of machine translation on the performance of Portuguese language models, multilingual models, and models in languages similar to Portuguese.*

We achieve this objective by translating the selected datasets into English, Spanish and Galician, and then evaluating the performance of the models on translated datasets. One of the questions we seek to answer through these experiments is whether models pretrained on languages similar to Portuguese can outperform models pretrained on general-purpose Portuguese data if they are fine-tuned on translated datasets.

1.3 | Proposed Solution

Our proposed solution involves the creation of the **Na**tural **Po**rtuguese **La**nguage **B**enchmark [1]. This benchmark comprises Portuguese datasets selected based on criteria sourced from literature on benchmark design principles Wang et al. (2019). We extract raw datasets from prior works, adapting them into a format suitable for language model evaluation, with some being used for this purpose for the first time.

Moreover, we ensure a comprehensive selection of language models that encompasses the most proficient ones for Portuguese natural language processing. This includes multilingual models pretrained on substantial amounts of Portuguese data. Notably, some models, like DeBERTa v3 He et al. (2023), are evaluated on Portuguese natural language tasks within an academic context for the first time in this study.

Furthermore, we've established an experimental framework that permits rigorous evaluation of the chosen language models on the benchmark. This ensures a fair comparison, producing results that genuinely reflect the models' performance. Each model undergoes testing on each task under diverse hyperparameter settings, making certain that factors such as random seed choice and hyperparameter initialization don't bias our comparisons.

Following Fu et al. (2020), we don't merely conduct a general evaluation; we also delve into a fine-grained assessment complemented by thorough statistical analysis of the outcomes. This approach reveals facets of our evaluation that might remain obscured in a more generic assessment.

Lastly, addressing our third research objective, we experiment with machine translation by converting the selected datasets into English, Spanish, and Galician.

[1] `https://github.com/ruanchaves/napolab`

This provides valuable insights into the efficacy of machine translation within the evaluation context of the chosen language models.

1.4 | Document Structure

The structure of this document is as follows:

1.4.1 | Background and Literature Overview

This chapter provides a thorough examination of developments in Portuguese Natural Language Processing (NLP), highlighting datasets, benchmarks, and evaluation methods specifically tailored for Portuguese language models. We introduce datasets that serve a range of applications, from semantic similarity to hate speech detection, and emphasize the unique linguistic challenges faced when constructing Portuguese benchmarks. The evolution of Portuguese language models is traced from the early statistical methods to the present neural network-driven era, with a particular focus on Portuguese and multilingual language models pertinent to our study. Lastly, this chapter deliberates over the significance of robust evaluation techniques in NLP, offering an overview of fine-grained analysis and statistical significance testing.

1.4.2 | Materials and Methods

This chapter delves into the design and implementation of an experimental system for evaluating pre-trained language models on Portuguese natural language tasks. It covers the system's layout, data gathering and pre-processing with a focus on dataset selection and modification. An intricate methodology was adopted to create train-validation-test splits, reducing translation artifacts and ensuring dataset consistency. A GPU Cluster Pipeline, consisting of a hyperparameter search, random seed selection, and final fine-tuning, was developed to ensure a fair comparison between the selected models. A nuanced evaluation approach for classification tasks was introduced, encompassing a fine-grained evaluation supplemented by proper statistical tests. The chapter thus serves as a precursor to the analysis of results in the next chapter.

1.4.3 | Results and Discussion

This chapter offers an exhaustive review of experimental results, detailed model evaluations, and statistical tests to ensure the findings' reliability. Through the use of the Friedman and Almost Stochastic Order (ASO) tests, it was determined that there's no significant performance difference between models trained on Portuguese-only data versus multilingual data. The ASO test revealed no single model dominated universally, while a fine-grained evaluation found mDeBERTa He et al. (2023) had

superior performance in handling out-of-vocabulary words on the ASSIN datasets. By comparing findings to previous studies, the chapter emphasizes the importance of detailed model evaluation, the potential benefits of language-specific tokenizers, and the necessity of varied tasks for comprehensive model evaluation. The efficacy of training models on specific Portuguese variants is also questioned, suggesting further exploration into tasks sensitive to variant-specific linguistic nuances.

1.4.4 | Conclusion

In this chapter, we revisit the steps taken and the results achieved in the evaluation of Portuguese language models. While we gathered valuable insights, future work aims to refine our evaluation methods and resources. In conclusion, we emphasize the need for robust benchmarks in the NLP community. We encourage researchers to move beyond SOTA-chasing Church and Kordoni (2022) and to place a stronger focus on real-world applicability.

2

Background and Literature Overview

This chapter provides an in-depth examination of existing research and developments in the field of Portuguese natural language processing. It sets the stage by discussing various datasets and benchmarks, shedding light on the principles that guide their design and the challenges in the evaluation of Portuguese language models.

2.1 | Portuguese Datasets

This section explores the different datasets available for the Portuguese language, while approaching the considerations behind benchmark design, and the selection criteria to ensure that appropriate datasets are selected for the evaluation of systems for Natural Language Understanding.

2.1.1 | Benchmark Design

While there are many datasets available for the Portuguese language to evaluate Natural Language Processing (NLP) systems, there's still little discussion about what makes a good benchmark for the evaluation of Portuguese language models, the principles that should guide its design, and the selection criteria for datasets. During the elaboration of the SuperGLUE benchmark, the work by Wang et al. (2019) has provided useful criteria to guide the creation of natural language understanding benchmarks:

- **Task substance**: Tasks should test a system's ability to understand texts in the target language.
- **Task difficulty**: Tasks should be challenging for current state-of-the-art systems but still solvable by educated speakers of the target language.

- **Evaluability**: Tasks must include an automatic metric that agrees well with human judgments of the output quality.
- **Public data**: Tasks should use publicly available training data, meaning it should be easily accessible to anyone.
- **Task format**: Tasks should have simple input and output formats to avoid overly complex, task-specific designs.
- **License**: Task data should be licensed for use and distribution in academic research.

For languages other than English, three additional considerations must be taken into account. These considerations haven't been discussed in the elaboration of the superGLUE benchmark, but they become important for languages with fewer resources.

2.1.1.1 | Translationese

The first consideration is the limited availability of datasets for languages like Portuguese. As a result, there is a common approach to use datasets translated from English for evaluating NLP systems. However, as highlighted by Artetxe et al. (2020), such datasets inherently gravitate towards a unique linguistic register, colloquially termed *translationese*. This register can differ significantly from everyday language usage, particularly in industrial NLP applications. Therefore, relying on these translated datasets may not provide fully accurate evaluation results.

2.1.1.2 | Translation Artifacts

The second consideration involves what are called *translation artifacts*. An example of this is seen in the PLUE benchmark, where the original GLUE benchmark was translated using Google Cloud translation by Gomes (2020). This translated benchmark has been used by Rodrigues et al. (2023) to evaluate the Albertina models.

A notable issue in this benchmark is that the sentence pairs were translated separately, including those for natural language inference tasks, rather than translated together as a single input. Artetxe et al. (2020) showed that translating the sentences separately can reduce the lexical overlap between them, leading to the introduction of translation artifacts. They also demonstrated that the performance of current language models is closely tied to this lexical overlap, and that therefore, the way the sentences are translated can significantly affect their performance.

2.1.1.3 | Automatic Data Annotation

A final consideration relates to the usage of automatic data annotation techniques. While this approach can quickly generate large datasets, it poses concerns regard-

ing label accuracy. Without human review, these labels might introduce biases and inaccuracies. As a case in point, the "B2W-Reviews" dataset by Real et al. (2019), which consists of Portuguese customer product reviews, contains star ratings that haven't been validated by humans. Bigne et al. (2023) and Mudambi et al. (2014) suggest that these automated ratings might not always match the actual sentiment of the review, underscoring the need for human verification or improved annotation techniques.

2.1.2 | Selected Datasets

This section provides an overview of the selected Portuguese datasets, curated in alignment with the criteria detailed in Section 2.1.1. The datasets were selected based on their pertinence to the study and their compliance with the specific requisites of our analysis.

The rationale behind our selection is expounded in Table 2.1. As delineated in the table, existing Portuguese benchmarks did not satisfy our criteria, necessitating the compilation of our own set of suitable datasets. For instance, Pires et al. (2023) introduced Poeta, a benchmark composed of 14 Portuguese datasets, of which half are translations from English. Among the native datasets, TweetSentBR Brum and Nunes (2017) is accessible only through the Twitter API. Therefore, some tweets in the dataset may become inaccessible due to deletion by the author or the platform. Indeed, Zubiaga (2018) discovered that the availability of tweets and unique users in a Twitter dataset can decline below 70% within four years post-publication.

Table 2.1 outlines the criteria employed for dataset selection, which are grounded in the related work discussion on benchmark design from Section 2.1.1. The criteria are as follows:

- **Public**: A dataset should be readily available for acquisition through a public link, obviating the necessity for intricate proprietary API interactions (e.g., the Twitter API) or the completion of request forms. Notably, the Poeta benchmark does not satisfy this criterion due to the unavailability of some of its datasets, including TweetSentBR Brum and Nunes (2017). Similarly, other datasets based on Twitter data, including TwitterDialogueSAPT Carvalho et al. (2022) and the dataset proposed by Peres et al. (2017), are not publicly accessible. Datasets affiliated with the IberLEF conferences, including IDPT 2021 and ABSAPT 2022, are also not available in the public domain, necessitating direct communication with the conference organizers for acquisition Corrêa et al. (2021); da Silva et al. (2022).

- **Reliable**: A dataset should employ an evaluation metric that manifests a strong and consistent correlation with human judgments. Datasets that resort to metrics like BLEU or METEOR for evaluation are thereby rejected. Our survey did not identify any datasets that failed to meet this criterion.

- **Natural**: The datasets should encompass native Portuguese text or text that has undergone professional translation. The PLUE benchmark and the AIA-BDE dataset were excluded exclusively based on this criterion Gomes (2020); Gonçalo Oliveira et al. (2020). A review by professional translators would render these datasets eligible for our selection.
- **Human**: The tasks within the dataset should be annotated by human experts, ensuring that the labels are not auto-generated. Datasets such as B2W-Reviews01 and ClozeQuestions were excluded based solely on this criterion Oliveira et al. (2014); Real et al. (2019). A thorough review by human annotators would render these datasets eligible for inclusion.
- **General**: The tasks within the dataset should necessitate only general domain knowledge, removing the need for expertise that cannot be reasonably expected from an educated Portuguese speaker. Datasets such as LeNER-BR, which belongs to the legal domain, FakeWhatsappBR, which demands intricate knowledge of Brazilian politics, and DIP, which requires familiarity with Brazilian literary works, do not meet this criterion Cunha (2021); Luz de Araujo et al. (2018); Simões (2023).

Dataset	Public	Reliable	Natural	Human	General
Benchmarks					
PLUE Gomes (2020)	✓	✓	✗	✓	✓
Poeta Pires et al. (2023)	✗	✓	✗	✓	✗
Accepted Datasets					
ASSIN Fonseca et al. (2016)	✓	✓	✓	✓	✓
ASSIN 2 Real et al. (2020)	✓	✓	✓	✓	✓
PorSimplesSent Leal et al. (2018b)	✓	✓	✓	✓	✓
ReLi Freitas et al. (2014)	✓	✓	✓	✓	✓
HateBR Vargas et al. (2022)	✓	✓	✓	✓	✓
FaQUAD Sayama et al. (2019)	✓	✓	✓	✓	✓
HAREM Santos et al. (2006)	✓	✓	✓	✓	✓
ReRelEM Freitas et al. (2009)	✓	✓	✓	✓	✓
MacMorpho Aluísio et al. (2003)	✓	✓	✓	✓	✓
Rejected Datasets					
LeNER-BR Luz de Araujo et al. (2018)	✓	✓	✓	✓	✗
Peres et al. (2017)	✗	✓	✓	✓	✓
TweetSentBR Brum and Nunes (2017)	✗	✓	✓	✓	✓
B2W-Reviews01 Real et al. (2019)	✓	✓	✓	✗	✓
FakeWhatsappBR Cunha (2021)	✓	✓	✓	✓	✗
TwitterDialogueSAPT Carvalho et al. (2022)	✗	✓	✓	✓	✓
AIA-BDE Gonçalo Oliveira et al. (2020)	✓	✓	✗	✓	✓
ClozeQuestions Oliveira et al. (2014)	✓	✓	✓	✗	✓
ABSAPT 2022 da Silva et al. (2022)	✗	✓	✓	✓	✓
FACTCK BR Moreno and Bressan (2019)	✓	✓	✓	✓	✗
IDPT 2021 Corrêa et al. (2021)	✗	✓	✓	✗	✓
DIP Simões (2023)	✓	✓	✓	✗	✗

Table 2.1: We examine various Portuguese datasets and classify them as accepted or rejected, based on a set of criteria delineated in the columns of this table. **Public** signifies that a dataset is readily accessible for download through a public link. **Reliable** denotes that the tasks associated with the dataset employ a reliable evaluation metric. **Natural** implies that the datasets consist of native Portuguese text or text that has been professionally translated. **Human** indicates that the tasks have been annotated by humans. Lastly, **General** suggests that the tasks included in the dataset necessitate only general domain knowledge for completion.

We delve below into the selected datasets, given a brief summary of their characteristics and a few samples to illustrate their data. We strived to present the samples in format as close as possible to the original format provided by the authors of the datasets, in order to facilite later discussions in this study about how the datasets were transformed.

For all the datasets, we present both the original and translated samples. For the purpose of building these tables, the samples were automatically translated into English using Google Translate. It should be noted that the success of the translations to keep the original characteristics of the data varied among the datasets, with notable failures in the case of the HateBR (Section 2.1.2.6) and PorSimplesSent (Section 2.1.2.4) datasets.

2.1.2.1 | ASSIN

```
<pair entailment="Entailment" id="1" similarity="4.0">
    <t>Mau tempo abate-se sobre o Algarve e inunda a baixa de
       Albufeira.</t>
    <h>A baixa de Albufeira foi afetada pelo mau tempo.</h>
</pair>

<pair entailment="Paraphrase" id="2" similarity="5.0">
    <t>Em comparação com o ano anterior, registaram-se menos 29
       acidentes e menos duas vítimas mortais.</t>
    <h>Feita a comparação com igual período do ano passado,
       registaram-se menos 29 acidentes e menos dois mortos.</h>
</pair>

<pair entailment="None" id="3" similarity="1.75">
    <t>Presidente da República fez uma série de avisos ao novo
       primeiro-ministro na tomada de posse do XXI Governo
       Constitucional.</t>
    <h>Foi naquele local que António Costa tomou posse do XXI
       Governo Constitucional.</h>
</pair>
```

Figure 2.1: Examples of premises (<t> tag) and hypotheses (<h> tag) from the ASSIN dataset showcasing different types of entailment relationships: `Entailment`, `Paraphrase`, and `None`. Each pair is accompanied by its ID and a semantic similarity score between the premise and hypothesis, which ranges from 0 to 5.

```
<pair entailment="Entailment" id="1" similarity="4.0">
    <t>Bad weather hits the Algarve and floods downtown Albufeira.
       </t>
    <h>Downtown Albufeira was affected by bad weather.</h>
</pair>

<pair entailment="Paraphrase" id="2" similarity="5.0">
    <t>Compared to the previous year, there were 29 fewer
       accidents and two fewer fatalities.</t>
    <h>Compared with the same period last year, there were 29
       fewer accidents and two deaths.</h>
</pair>

<pair entailment="None" id="3" similarity="1.75">
    <t>The President of the Republic issued a series of warnings
       to the new Prime Minister on the inauguration of the XXI
       Constitutional Government.</t>
    <h>It was there that António Costa took office for the XXI
       Constitutional Government.</h>
</pair>
```

Figure 2.2: Examples of premise and hypothesis pairs from Figure 2.1 after automatic translation from Portuguese to English.

ASSIN Fonseca et al. (2016) is a dataset comprising sentence pairs annotated for both semantic similarity scores and entailment. Each sentence pair is assigned a similarity score along with a label, which can be either `Entailment`, `Paraphrase`, or `None` to indicate the absence of both entailment and paraphrase relationships between the sentences.

The ASSIN dataset is also notable for having separate training and evaluation splits for European and Brazilian Portuguese. A total of 10,000 sentence pairs were sourced from Google News, with an equal division between Brazilian and European Portuguese sources.

Figures 2.1 and 2.4 showcase three sentence pairs extracted from the European Portuguese test set of ASSIN. Given that the content was extracted from news sources, there is a notable prevalence of named entities throughout the dataset. As evidenced in the figures, there are mentions of numerical data, references to geographic regions in Portugal (e.g., Algarve, Albufeira), and named entities in Portuguese politics, such as António Costa.

2.1.2.2 | ASSIN 2

```
<pair entailment="None" id="0" similarity="3.8">
    <t>O cachorro caramelo está assistindo um cachorro castanho
       que está nadando em uma lagoa</t>
    <h>Um cachorro de estimação está de pé no banco e está olhando
        outro cachorro, que é castanho, na lagoa</h>
</pair>

<pair entailment="Entailment" id="1" similarity="3.75">
    <t>O cara está fazendo exercícios no chão</t>
    <h>Um cara está fazendo exercícios</h>
</pair>

<pair entailment="Entailment" id="2" similarity="4.4">
    <t>Um cachorro grande e um cachorro pequenino estão parados ao
        lado do balcão da cozinha e estão investigando</t>
    <h>Um cachorro grande e um cachorro pequenino estão de pé no
       balcão da cozinha e investigam</h>
</pair>
```

Figure 2.3: Examples of premises (<t> tag) and hypotheses (<h> tag) from the ASSIN 2 dataset showcasing different types of entailment relationships: `Entailment` and `None`. Each pair is accompanied by its ID and a semantic similarity score between the premise and hypothesis, which ranges from 0 to 5.

ASSIN 2 Real et al. (2020) follows the same premise as the ASSIN dataset but with a few modifications. The `Paraphrase` category was removed, and the entailment annotations were limited to either `Entailment` or `None`. Another significant change concerns the source of the sentences: part of the dataset was translated from English, specifically from the SICK benchmark Marelli et al. (2014), and then manually corrected and annotated by professional linguists. The remaining sentences were composed in the same vein as the SICK benchmark, aiming to maintain simplicity in vocabulary and sentence structure.

This dataset does not distinguish between Brazilian and European Portuguese. While most sentences are notably straightforward and could apply to either variant, there's a noticeable lean towards Brazilian Portuguese in the dataset.

Figures 2.1 and 2.4 showcase three sentence pairs from the test set of ASSIN 2. It's evident that the linguistic context of the dataset is limited to basic, everyday life scenarios. This aligns with the dataset's intent to evaluate language understanding capabilities of models rather than the depth of their in-domain cultural knowledge

```
<pair entailment="None" id="0" similarity="3.8">
    <t>The caramel dog is watching a brown dog that is swimming in
        a pond</t>
    <h>A pet dog is standing on the bench and is looking at
       another dog, who is brown, in the pond</h>
</pair>

<pair entailment="Entailment" id="1" similarity="3.75">
    <t>The guy is doing exercises on the floor</t>
    <h>A guy is doing exercises</h>
</pair>

<pair entailment="Entailment" id="2" similarity="4.4">
    <t>A big dog and a small dog are standing next to the kitchen
       counter and are investigating</t>
    <h>A big dog and a small dog are standing on the kitchen
       counter and investigating</h>
</pair>
```

Figure 2.4: Examples of premise and hypothesis pairs from Figure 2.3 after automatic translation from Portuguese to English.

associated with Portuguese-speaking communities.

2.1.2.3 | ReRelEM

ReRelEM Freitas et al. (2009) is a dataset designed for extracting relations between named entities. It is built upon the HAREM dataset Freitas et al. (2010); Mota and Santos (2008); Santos et al. (2006), which was originally designed for named entity recognition. The documents were collected for annotation from a wide variety of internet sources, such as Wikipedia, blogs, and newspapers. The HAREM dataset had relationships between its entities annotated with 24 relation types.

As shown in Figures 2.5 and 2.6, ReRelEM includes additional annotations beyond entity relations. Relations are annotated in the `COREL` field of each entity, while the types of these relations are indicated in the `TIPOREL` field. When multiple relations are annotated, the n-th tag in the `COREL` field corresponds to the n-th tag in the `TIPOREL` field. Other XML fields provide additional details about the entities, such as their category (`CATEG`), type (`TIPO`) and subtype (`SUBTIPO`).

2.1.2.4 | PorSimplesSent

```
<DOC DOCID="H2-dftre765">
<P>
    A imprensa, inventada na

    <EM ID="H2-dftre765-9" CATEG="LOCAL" TIPO="HUMANO"
       SUBTIPO="PAIS" COREL="H2-dftre765-37" TIPOREL="
       incluido">Alemanha</EM>

    por

    <EM ID="H2-dftre765-10" CATEG="PESSOA" TIPO="
       INDIVIDUAL" COREL="H2-dftre765-9" TIPOREL="
       natural_de">John Gutenberg</EM>

  , foi importante na divulgação destas ideias. As

  <ALT>
  <EM ID="H2-dftre765-12aa" CATEG="OBRA" TIPO="REPRODUZIDA
      ">95 Teses de Martinho Lutero</EM> |
  <EM ID="H2-dftre765-12" CATEG="OBRA" TIPO="REPRODUZIDA"
      SUBTIPO="LIVRO">95 Teses</EM>
  de
  <EM ID="H2-dftre765-13" CATEG="PESSOA" TIPO="INDIVIDUAL"
       COREL="H2-dftre765-12␣H2-dftre765-9␣H2-dftre765-1"
      TIPOREL="autor_de␣natural_de␣PESSOA**participante_em
      **H2-dftre765-1**ACONTECIMENTO">Martinho Lutero</EM>
      </ALT>
</P>
</DOC>
```

Figure 2.5: An XML snippet from the ReRelEM dataset shows annotated entities and relationships. This excerpt from the document with ID `H2-dftre765` references the invention of the press in Germany by Johannes Gutenberg and emphasizes Martin Luther's "95 Theses". It's evident that `Martinho Lutero` (Portuguese for "Martin Luther"), the last entity in the snippet, has a relation (indicated by the `COREL` attribute) with the entity having ID `H2-dftre765-12`, which contains the text `95 Teses` (Portuguese for "95 Theses"). The relationship of Martin Luther with the 95 Theses is denoted as `autor_de` (Portuguese for "author of") in the `TIPOREL` attribute.

```
<DOC DOCID="H2-dftre765">
<P>
    The press, invented in

    <EM ID="H2-dftre765-9" CATEG="PLACE" TYPE="HUMAN" SUBTYPE="
       COUNTRY" COREL="H2-dftre765-37" RELTYPE="included">Germany
       </EM>

    by

    <EM ID="H2-dftre765-10" CATEG="PERSON" TYPE="INDIVIDUAL" COREL
       ="H2-dftre765-9" RELTYPE="from">John Gutenberg</EM>

    , was crucial in spreading these ideas. The

    <ALT>
    <EM ID="H2-dftre765-12aa" CATEG="WORK" TYPE="REPRODUCED">95
       Theses of Martin Luther</EM> |
    <EM ID="H2-dftre765-12" CATEG="WORK" TYPE="REPRODUCED" SUBTYPE
       ="BOOK">95 Theses</EM>
    of
    <EM ID="H2-dftre765-13" CATEG="PERSON" TYPE="INDIVIDUAL" COREL
       ="H2-dftre765-12␣H2-dftre765-9␣H2-dftre765-1" RELTYPE="
       author_of␣from␣PERSON**participant_in**H2-dftre765-1**
       EVENT">Martin Luther</EM></ALT>
</P>
</DOC>
```

Figure 2.6: An XML snippet from the ReRelEM dataset from Figure 2.5 after automatic translation from Portuguese to English.

production_id	level	changed	split	sentence_text_from	sentence_text_to
3	NAT->STR	S	N	Ele baixou de posto no exército.	Ele foi condenado a baixar de posto no exército.
3	NAT->STR	N	N	Ele foi condenado ontem por um tribunal militar a 90 dias de trabalho forçado sem prisão.	Ele foi condenado ontem por um tribunal militar a 90 dias de trabalho forçado sem prisão.
42	ORI->NAT	S	N	A Casa Branca teria dado instruções à delegação americana em Bali para evitar que o país seja responsabilizado pelo fracasso das negociações.	A Casa Branca teria dado instruções à comissão de representantes americana em Bali para evitar que o país seja considerado responsável pelo fracasso das negociações.
42	ORI->NAT	S	N	As estratégias já começaram a envolver os altos escalões dos governos.	As estratégias já começaram a envolver os altos níveis dos governos.
165	NAT->STR	S	S	Smith tem um braço-robô com uma cobertura de pele artificial que pode ser reposta com um spray.	Smith tem um braço-robô com uma cobertura de pele artificial.
165	NAT->STR	N	N	A tecnologia de Someya e colegas ainda está muito longe disso.	A tecnologia de Someya e colegas ainda está muito longe disso.

Table 2.2: Sample sentence pairs from the PorSimplesSent dataset illustrate the simplification operations applied to original sentences from Brazilian newspapers. In each pair, both sentences convey the same meaning. The sentence on the right is either more readable than the one on the left or contains identical text. Any changes between the two sentences are highlighted in gray. The **level** column indicates the type of simplification operation. The **changed** column shows whether any modifications were made, and **split** denotes if the sentence on the left was divided into multiple sentences during the simplification process.

production_id	level	changed	split	sentence_text_from	sentence_text_to
3	NAT->STR	S	N	He dropped ranks in the army.	He was sentenced to downgrade in the army.
3	NAT->STR	N	N	He was sentenced yesterday by a military court to 90 days of forced labor without arrest.	He was sentenced yesterday by a military court to 90 days hard labor without arrest.
42	ORI->NAT	S	N	The White House would have given instructions to the American delegation in Bali to avoid the country being held responsible for the failure of the negotiations.	The White House would have given instructions to the commission of American representatives in Bali to prevent the country from being considered responsible for the failure of the negotiations.
42	ORI->NAT	S	N	Strategies have already started to involve the highest levels of government.	Strategies have already started to involve the highest levels of government.
165	NAT->STR	S	S	Smith has a robot arm with an artificial skin covering that can be replaced with a spray.	Smith has a robot arm with an artificial skin covering.
165	NAT->STR	N	N	The technology of Someya and colleagues is still a long way from that.	Someya and colleagues' technology is still a long way from that.

Table 2.3: Sample sentence pairs from the PorSimplesSent dataset (see Table 2.2) translated automatically from Portuguese to English. Any changes between the two sentences are highlighted in gray. It should be noted that the automatic translation did not preserve the original simplification levels of each sentence.

PorSimplesSent Leal et al. (2018a) is a dataset for the task of sentence readability assessment. Each sentence pair in the dataset conveys the same basic meaning; however, one sentence in the pair is more readable than the other.

All sentences in the dataset were collected from Brazilian newspapers. They were then manually simplified by human annotators in two distinct ways: either through natural simplification or strong simplification, with the latter being less conservative about what is retained in the sentence.

Figure 2.2 displays the structure of the sentence pairs in the PorSimplesSent dataset.

The dataset contains the following fields:

- **production_id**: The ID of the sentence pair.
- **level**: This indicates the kind of operation that was performed to transform the left sentence into the right one. `NAT->STR` indicates that a sentence that underwent natural simplification was converted into a strongly simplified version. `ORI->NAT` indicates that an original sentence was turned into a naturally simplified version.
- **changed**: A field indicating whether or not any changes were made to the sentence during the simplification process. If changes were made, this field is marked as "S" ("Sim", Portuguese for "Yes"). If no changes were made and both sentences in the pair are identical, the field is marked as "N" ("Não", Portuguese for "No").
- **split**: Indicates whether the simplification process resulted in splitting a single sentence into multiple sentences. Marked "S" if true and "N" if false.
- **sentence_text_from**: The left sentence of the pair.
- **sentence_text_to**: The right sentence of the pair.

Figure 2.3 showcases samples from PorSimplesSent after automatic translation. It's crucial to highlight that, due to the subtle nature of this task, automatic translation may not preserve the annotations' consistency in the dataset. In the original dataset, the right sentence is simpler than the left one. However, post-translation, this relationship can be inverted or its nature might change, e.g., a natural simplification might become a strong simplification, and vice versa.

2.1.2.5 | ReLi

ReLi Freitas et al. (2014) is a dataset for fine-grained sentiment analysis. The dataset consists of customer reviews of popular literary works written by seven distinct authors. The dataset contains a total of 1600 reviews, each annotated for sentiment

word	pos	object	opinion	polarity	help
Foi	V	O	O	+	O
um	ART	O	O	+	O
boa	ADJ	O	op00+	+	O
leitura	N	O	op00+	+	O
,	,	O	O	+	O
mas	KC	O	O	+	O
nada	ADV	O	O	+	O
excepcional	ADJ	O	op00-	+	O
.	.	O	op00-	+	O
"	"	O	O	+	HELP
O	ART	OBJ00	O	+	HELP
Apanhador	N	OBJ00	O	+	HELP
em	PREP	OBJ00	O	+	HELP
o	ART	OBJ00	O	+	HELP
Campo	NPROP	OBJ00	O	+	HELP
de	NPROP	OBJ00	O	+	HELP
Centeio	NPROP	OBJ00	O	+	HELP
"	"	O	O	+	HELP

Table 2.4: Two sentence samples extracted from book reviews on the ReLi dataset. The **pos** column provides part of speech tags for the words in the review. The **object** column contains the entity ID of the entity to which the word is referring to. **opinion** indicates that the word expresses an opinion about an entity with the given ID. **polarity** indicates the local sentiment polarity of the word, and **help** indicates if annotators judged the sentence hard to annotate and had to request help to decide on it.

polarity (positive or negative) at the token level. Each word has its own sentiment annotation.

The dataset also provides part of speech annotations and supplementary annotations to indicate the entity to which a sentence refers, and the sentiment of the sentence towards that entity. Each review is also accompanied by a star rating given by the customer.

As indicated in Table 2.4, the columns in the ReLi dataset are as follows:

- **word**: the words in the book review. Table 2.4 displays two sentences extracted from book reviews in the ReLi dataset: *"Foi uma boa leitura, mas nada de excepcional"* (translated as *"It was a nice read, but nothing exceptional"*), and *"O Apanhador em o Campo de Centeio"* (translated as *"The Catcher in the Rye"*).

- **pos**: part of speech tags for the words in the book review, manually annotated by linguists.

- **object**: the ID of the object to which the sentence refers. In the given example in Table 2.4, the book "The Catcher in the Rye" is the referred object with ID OBJ00.
- **opinion**: the polarity of the opinion about a given object, accompanied by the ID of the referred object. For instance, in Table 2.4, "op00-" indicates a negative opinion about the object with ID OBJ00.
- **polarity**: The local sentiment polarity of the word, considered in its immediate sentence context.
- **help**: The difficulty level the annotator experienced when annotating this sentence. "HELP" indicates that they requested assistance to complete the annotation.

2.1.2.6 | HateBR

instagram_comments	offensive_language	offensiveness_levels	hate_speech
Essa nao tem vergonha na cara!!	1	2	-1
"Quem tem pena é galinha, mas ela é uma VACA LOUCA."	1	3	8
Oportunista essa corrupta.Agora todos os Comunistas querem se fazer de vítimas.	1	3	5

Table 2.5: Examples of Instagram comments from the HateBR dataset showcasing annotations for offensive language (**offensive_language**), degree of offensiveness (**offensiveness_levels**), and specific categories of hate speech (**hate_speech**). Comments can be classified as offensive without necessarily being labeled as hate speech. Levels of offensiveness are indicated in the column **offensiveness_levels**) and can range from 0 (Non-offensive) to 3 (Highly offensive), while specific hate speech categories are assigned unique IDs under the **hate_speech** column, such as sexism (8) and partyism (5).

HateBR Vargas et al. (2022) is a dataset constructed from seven thousand Instagram comments on profiles of Brazilian politicians for the detection of offensive language and hate speech. Each comment has been annotated by experts using both a binary classification (offensive vs. non-offensive) and a fine-grained hate speech classification. The dataset contains the following attributes, as indicated in Tables 2.5 and 2.6:

instagram_comments	offensive_language	offensiveness_levels	hate_speech
This one has no shame in her face!!	1	2	-1
"It's a chicken who feels sorry for her, but she's a MAD COW."	1	3	8
This corrupt opportunist. Now all Communists want to play victims.	1	3	5

Table 2.6: Sample sentences from the HateBR dataset (see Table 2.5) translated from Portuguese to English. Due to the idiomatic nature of the original Instagram comments, the automatic translation may not capture all nuances in meaning.

- **instagram_comments**: The text of the Instagram comment.
- **offensive_language**: This attribute is a binary classification of the comment as either offensive (1) or non-offensive (0). Note that a comment can be deemed offensive without being classified as hate speech. For instance, a comment might contain swear words but not exhibit content that qualifies as hate speech.
- **offensiveness_levels**: This attribute indicates the degree of offensiveness: Highly offensive (3), Moderately offensive (2), Slightly offensive (1), Non-offensive (0).
- **hate_speech**: Tables 2.5 and 2.6 provide examples of comments classified as offensive but not hate speech (-1), sexism (8), and partyism (5). The fine-grained categories of hate speech, along with their respective IDs, are as follows: Antisemitism: 1, Apology for the dictatorship: 2, Fatphobia: 3, Homophobia: 4, Partyism: 5, Racism: 6, Religious intolerance: 7, Sexism: 8, Xenophobia: 9, Offensive but not hate speech: -1, Non-offensive: 0.

The linguistic register of the HateBR dataset is highly casual and informal, as reflected in its heavy use of swear words and idiomatic expressions. It can be seen in Table 2.6 that automatic translation systems, designed primarily for a more neutral form of the Portuguese language, struggle to preserve all the nuances of the original comments.

2.1.2.7 | FaQuaD

FaQuaD Sayama et al. (2019) is a reading comprehension dataset that follows the format of SQuAD Rajpurkar et al. (2016). It consists of 900 reading comprehen-

sion questions about 249 paragraphs taken from documents and Wikipedia articles related to Brazilian higher education institutions.

To provide a clearer understanding of the structure of the dataset, Figures 2.7 and 2.8 depict a sample JSON entry from the FaQuaD dataset. The JSON object contains a "title" that represents the name and unique identifier of the document, which includes a set of paragraphs. The "paragraphs" array includes the context, that is, the text of the paragraph, and the associated "qas" (questions and answers). Within "qas", each entry has a "question", its "id", and either a single or multiple potential "answers". Each "answer" contains the textual response and its starting character index in the context. The answers are, therefore, purely extractive and refer to specific spans in the given context.

```
{
  "title": "ENGENHARIA_DE_COMPUTACAO",
  "paragraphs": [
    {
      "context": "Atualmente, uma nova metodologia de ensino vem
         sido aplicada a cursos de Engenharia de Computação no
         Brasil: a Aprendizagem baseada em problemas ou Problem
         Based Learning (PBL). Essa metodologia, que tem como
         principal objetivo explorar o autodidatismo do aluno,
         bem como sua capacidade de trabalho em grupo, já vem
         sendo aplicada em cursos novos de EC no Brasil, como na
         Universidade Estadual de Feira de Santana (UEFS), que
         foi a pioneira no Brasil, e na Universidade Federal da
         Bahia (UFBA).",
      "qas": [
        {
          "question": "Qual metodologia vem sendo aplicada a
             cursos de engenharia de computação no Brasil?",
          "id": "9f246d4cc95f4ae7a1831f54443bb1eb",
          "answers": [
            {
              "text": "Aprendizagem baseada em problemas",
              "answer_start": 111
            },
            {
              "text": "Problem Based Learning (PBL)",
              "answer_start": 148
            }
          ]
        }
      ]
    }
  ]
}
```


Figure 2.7: A JSON sample from the FaQuaD dataset. The JSON object contains a "title" representing the name of the document, a "paragraphs" array that includes the context and associated "qas" (questions and answers). Within "qas", each entry has a "question", its "id", and potential "answers", each of which contains the textual answer and its starting character index in the context. This example showcases a paragraph discussing the application of Problem Based Learning (PBL) in Brazilian Computer Engineering courses.

```
{
  "title": "COMPUTER_ENGINEERING",
  "paragraphs": [
    {
      "context": "Currently, a new teaching methodology has
         been applied to Computer Engineering courses in
         Brazil: Problem-Based Learning (PBL). This
         methodology, which aims mainly to explore the
         student's self-taught abilities as well as their
         group work skills, has already been implemented in
         new EC courses in Brazil, such as at the State
         University of Feira de Santana (UEFS), which was the
          pioneer in Brazil, and the Federal University of
         Bahia (UFBA).",
      "qas": [
        {
          "question": "Which methodology has been applied to
             computer engineering courses in Brazil?",
          "id": "9f246d4cc95f4ae7a1831f54443bb1eb",
          "answers": [
            {
              "text": "Problem-Based Learning",
              "answer_start": 111
            },
            {
              "text": "Problem Based Learning (PBL)",
              "answer_start": 148
            }
          ]
        }
      ]
    }
  ]
}
```


Figure 2.8: A JSON sample from Figure 2.7 after automatic translation from Portuguese to English.

2.2 | Portuguese Language Models: A Historical Overview

This section seeks to delve into the chronological development and critical shifts in the application of NLP to Portuguese. Patil et al. (2023) highlights the introduction of neural network-based techniques with static word embeddings, where both syntactic and semantic features are learned by a neural network. This shift from statistical language models (STMs) to neural language models (NLMs) is also noted by Zhao et al. (2023). The section will begin with this pivotal transition.

2.2.1 | Neural Language Models

According to Zhao et al. (2023), utilizing language models for representation learning, rather than mere word sequence modeling, started a new stage in NLP. Neural language models such as word2vec by Mikolov et al. (2013) enabled the creation of distributed word representations.

2.2.1.1 | Portuguese static word embeddings

The first static word embeddings specifically trained for the Portuguese language and made publicly available were created by Rodrigues et al. (2016) for both European and Brazilian Portuguese variants, using a corpus collected from newspapers comprising 1.7 billion tokens.

Subsequently, Hartmann et al. (2017) trained static word embedding architectures on a corpus with 1.3 billion tokens, augmenting a filtered version of the training data used by Rodrigues et al. (2016) with content from Wikipedia, newspapers, magazines, and books.

In 2018, the BrWaC corpus was released by Wagner Filho et al. (2018). Since its release, language models for Portuguese have predominantly used this dataset as part of their training data. This corpus contains 2.68 billion tokens and was crawled from *.br* internet domains, ensuring that it consists almost exclusively of the Brazilian variant of Portuguese.

Santos et al. (2020) trained word embeddings on BrWaC, a corpus of blog posts, and a Wikipedia dump from 2019, totaling 4.9 billion tokens. Their work includes a comparison of the performance of their embeddings with those trained by Hartmann et al. (2017) on downstream tasks using the HAREM and ASSIN datasets. They concluded that their embeddings achieved inferior performance, as BrWaC had a lower domain coverage and diversity compared to the training data assembled by Hartmann et al. (2017).

2.2.1.2 | Variant-specific studies

At this juncture, the first studies emerged attempting to measure the effect of training neural language models on distinct variants of Portuguese, particularly comparing the outcomes of training them separately or concurrently. Although the European Portuguese and Brazilian Portuguese variants are mutually intelligible to some extent, significant syntactic and semantic differences between them may influence the performance of language models on downstream tasks. Fonseca and Aluísio (2016) identified a measurable degradation in performance when using word embeddings trained for one Portuguese variant to tag parts of speech in texts of another variant.

2.2.2 | Pre-trained Language Models

The development and utilization of Portuguese pre-trained language models (PLMs) include feature-based PLMs, fine-tuning-based PLMs, and Large Language Models (LLMs). By examining different models such as ELMo, BERTimbau, Albertina, and others, we offer a comprehensive overview of the development of pre-trained language models for the Portuguese language.

2.2.2.1 | Feature-based PLMs

ELMo ELMo, introduced by Peters et al. (2018), marked a significant advancement in the field of pre-trained language models within NLP. As highlighted by Devlin et al. (2018), ELMo's approach utilized a feature-based methodology, enabling task-specific architectures to enrich their models by integrating pre-trained ELMo representations as supplementary features. A distinctive aspect of ELMo is the implementation of contextual word embeddings, an innovative form of word representation that, contrary to static word embeddings, generates varying vectors for identical words depending on the surrounding context.

de Castro et al. (2019) trained two ELMo models as the first contextual word embeddings for the Portuguese language. One model was trained on BrWaC, and the other on a Wikipedia dump. These models were applied in the named entity recognition (NER) track at IberLEF 2019 Collovini et al. (2019), encompassing both general-purpose NER and specialized domain-specific tasks. The general-purpose dataset comprised the Second HAREM Freitas et al. (2010) and the SIGARRA datasets.

Comparison with other language models An evaluation conducted by Rodrigues et al. (2020) on the semantic similarity tasks using the ASSIN dataset revealed that the ELMo models trained by de Castro et al. (2019) were not only superior to static word embeddings but also outperformed embeddings produced by subsequent BERT models as pre-trained by Souza et al. (2020). Further support-

ing these findings, Hartmann and Aluísio (2020) confirmed ELMo's superiority over BERT for text simplification tasks within the SIMPLEX-PB-3.0 dataset.

It should be noted that the studies by Rodrigues et al. (2020) and Hartmann and Aluísio (2020) generated embeddings from BERT models by averaging the output layers, a method reported by Reimers and Gurevych (2019) to yield sub-optimal results in comparison to static word embeddings. They proposed a workaround to bring BERT to competitive performance by employing siamese networks to derive semantically meaningful sentence embeddings from BERT.

2.2.2.2 | Fine-tuning-based PLMs

The fine-tuning approach became popular with the introduction of masked language models by Devlin et al. (2018). Following this seminal work, the NLP community has developed numerous general-purpose pre-trained language models, as well as specialized fine-tuned versions for specific downstream applications.

Portuguese Language Models Approximately 50 general-purpose pre-trained language models are available on the Hugging Face Model Hub for the Portuguese language, six of which are intended exclusively for Portuguese NLP tasks.

In this section, we choose to highlight only the few general-purpose Portuguese models that have been published in peer-reviewed conferences or journals. This results in a concise list of models that have been widely adopted by the research community and that have been trained with sufficient computational resources to ensure competitive performance on various downstream tasks.

BERTimbau Souza et al. (2020) released *BERTimbau base* and *BERTimbau large*. Prior to pre-training, the weights of both models were initialized from BERT models published by Devlin (2020), except for the first and last layers (the embeddings layer and the masked language modeling head), which were assigned random weights.

BERTimbau base had its weights initialized from BERT Multilingual, while *BERTimbau large* was initialized from BERT Large. According to Devlin (2020), BERT Multilingual was pretrained on data from the Portuguese Wikipedia and other languages, and BERT Large was pretrained on the English Wikipedia and the BookCorpus dataset, released by Zhu et al. (2015). Bandy and Vincent (2021) later raised concerns about regarding the BookCorpus dataset, pointing out that around 35% of the books in the dataset are duplicates, and that book genres appear proportionate to their popularity, resulting in an overrepresentation of profanity and explicit sexual content.

Both BERTimbau models were pre-trained exclusively on the BrWaC dataset and evaluated by Souza et al. (2020) on the ASSIN 2 and HAREM datasets. Over the years, BERTimbau has become one of the fundamental models in Portuguese NLP,

frequently appearing in publications focused on various Portuguese-specific downstream tasks.

PTT5 PTT5, a T5 model pre-trained exclusively on the BrWaC dataset, was introduced by Carmo et al. (2020). Pre-training began from the T5 checkpoint released by Raffel et al. (2020), which was itself pre-trained on English data filtered from Common Crawl dumps with duplication and inappropriate content removed.

PTT5 was evaluated by Carmo et al. (2020) on the ASSIN 2 and HAREM datasets, registering performance inferior to BERTimbau. Nevertheless, it later surpassed BERTimbau in an evaluation by Gomes et al. (2022) on aspect-based sentiment analysis tasks.

Model	Dataset	Tokens	Documents
mDeBERTa XLM-RoBERTa	C100	23.875×10^9	37.305×10^6
BERTimbau Albertina-PTBR	BrWaC	2.68×10^9	3.5×10^6
Albertina-PTPT	OSCAR, DCEP, Europarl, ParlamentoPT	2.2×10^9	8×10^6
DeBERTa v2 DeBERTa v3	CC-News	2×10^9 $*$	2×10^6 $*$

Table 2.7: Corpora used for pretraining Portuguese and multilingual language models, together with the count of tokens and documents. An estimate ($*$) is given instead of the precise count of tokens and documents for the CC-News dataset because the exact size of the Portuguese subset of this corpus was not provided by the authors.

Albertina Rodrigues et al. (2023) released one model for each one of the main variants of the Portuguese language: Albertina-PTPT for European Portuguese and Albertina-PTBR for Brazilian Portuguese. Both were pre-trained starting from the same DeBERTa v2 checkpoint by He et al. (2020), retaining all weights, including the embeddings layer and masked language modeling head, and continuing the pre-training process with Portuguese data.

The DeBERTa v2 checkpoint was pre-trained on CC-News, containing a substantial amount of Portuguese data. As detailed in Table 2.7, the Portuguese subset of this dataset is comparable in scale to the widely-used BrWaC dataset.

Albertina-PTBR was pre-trained exclusively on BrWaC, while Albertina-PTPT utilized the OSCAR dataset, filtered for domains with the *.pt* extension, and documents from the European Parliament in European Portuguese. Albertina has been evaluated on the HAREM and ASSIN 2 datasets, but as of now, there are no further evaluations in the literature.

Multilingual models In alignment with Agerri and Agirre (2022), we also discuss two relevant multilingual models, XLM-RoBERTa and mDeBERTa, given their state-of-the-art results in Spanish natural language processing tasks, with a pre-training data proportion similar to the closely related Portuguese language.

mDeBERTa Multilingual DeBERTa was pre-trained by He et al. (2023) on the CC100 dataset, containing more than 23 billion tokens spread across approximately 37 documents in Brazilian Portuguese. Although the authors evaluated the model on a multilingual benchmark, Portuguese was not included in the evaluation.

mDeBERTa has outperformed BERTimbau in some studies. Gomes et al. (2022) reported that mDeBERTa achieved an F1 score of 85.5% on an Aspect Term Extraction task at the ABSAPT 2022 competition, compared to 82.6% by BERTimbau. In the Protest Document Classification task, organized by Hürriyetoğlu et al. (2022), mDeBERTa achieved an F1 score of 79.85%, compared to 77.96% by BERTimbau. Sahin et al. (2022) fine-tuned mDeBERTa on a dataset augmented with other languages, while BERTimbau was fine-tuned solely on Portuguese data.

XLM-RoBERTa *XLM-RoBERTa base* and *XLM-RoBERTa large* were released by Conneau et al. (2020) and trained on the CC100 dataset, without specific evaluation on Portuguese downstream tasks.

Vaidya and Kane (2023) reported inferior results using *XLM-RoBERTa base* instead of BERTimbau base in Portuguese dialect detection. However, Fernandes et al. (2022) found that *XLM-RoBERTa large* outperformed BERTimbau on the aforementioned Protest Document Classification task, although they also fine-tuned XLM-RoBERTa on an augmented multilingual dataset.

Models in other Iberian languages As this study includes machine translation experiments, we also cover monolingual models for languages similar to Portuguese, such as Spanish and Galician.

IXABERTesv2 A Spanish language model, IXABERTesv2 (IXAes) is based on RoBERTa base and trained on the OSCAR dataset. Agerri and Agirre (2022) found it to be the best monolingual option for Spanish, although still inferior to mDeBERTa and XLM-RoBERTa Large.

Bertinho Pre-trained by Vilares et al. (2021) from scratch on the Galician Wikipedia, Bertinho, a 12-layer BERT model, was reported to surpass BERT multilingual on selected Galician downstream tasks.

2.2.3 | Large Language Models

Zhao et al. (2023) defines a Large Language Model (LLM) as a Pretrained Language Model (PLM) that has been trained on a scale at which *emergent abilities* become possible. Although not every LLM may display such emergent abilities, a model is considered an LLM if it has a number of parameters for which these abilities have been documented as being possible. Such emergent abilities include, for instance, the ability to solve few-shot tasks through in-context learning.

Sabiá-65B Pires et al. (2023) pre-trained a LLaMA 65B model on the Portuguese subset of the ClueWeb22 dataset, which was released by Overwijk et al. (2022). ClueWeb22 is designed to be an improvement over CommonCrawl, aiming to provide a more realistic content distribution, enhance the quality of page content, and generate annotations for the page data. The model was evaluated on the Poeta benchmark, comprising both native and translated datasets, and achieved zero-shot results close to the state-of-the-art for supervised models. However, in evaluating LLMs with restrictive privacy policies regarding their training data, the risk of data contamination could not be ruled out. This refers to the possibility that the evaluation datasets themselves may have been included in the model's training data.

2.3 | Benchmarking Approaches in Pretrained Language Models

The evaluation and comparison of pretrained language models (PLMs) is a critical aspect in the field of Natural Language Processing. The proper benchmarking of these models has emerged as a subject of intense debate, as subtle variations in the experimental setup can lead to divergent conclusions. In this section, we delineate the principal considerations that must be addressed to ensure a comprehensive and unbiased comparison of PLMs.

2.3.1 | Hyperparameter Optimization Techniques

A widely observed concern in the benchmarking process pertains to the methods employed for hyperparameter tuning. While grid search is traditionally utilized for the evaluation of pretrained models in small-scale investigations, recent works have advocated for alternative approaches. Godbole et al. (2023) present compelling evidence in favor of quasi-random search for small-scale experiments, illustrating its superiority over grid search and Bayesian optimization methods.

Batch Size In light of findings by Shallue et al. (2019), Godbole et al. (2023) argue that adjusting the batch size should not be the primary means to enhance validation set performance. Instead, they demonstrate that aligning the tuning of optimizer

and regularization hyperparameters can achieve similar effects, and consequently, no compelling evidence supports the notion that batch size tuning contributes to superior validation performance.

Influence of Random Seed Initialization The choice of random seed initialization in the fine-tuning process is another fundamental consideration in the evaluation of PLMs. Dodge et al. (2020) have stated that a judicious selection of random seeds enables a BERT model to closely approach, and in some instances surpass, the performance of subsequently proposed models such as RoBERTa and XLNet.

The authors introduce a systematic approach for seed selection, termed "start many, stop early, continue some". This methodology starts numerous trials using distinct random seeds and proceeds to train only 30-50% of the most promising trials to completion, while prematurely halting the less promising ones at approximately 20-30% of the total training steps.

Fine-Tuning Dynamics and Overfitting Concerns The susceptibility to overfitting during fine-tuning of transformer-based masked language models has been rigorously investigated. Mosbach et al. (2020) provide a definitive analysis showing that overfitting does not present a significant issue, even when the number of iterations is augmented substantially, and the training loss is minimized nearly to zero.

Their work proposes that fine-tuning BERT for 20 epochs can be considered a robust baseline, thus challenging the assumption that early stopping is indispensable during the fine-tuning process. This insight is congruent with the results of Hao et al. (2019), who confirmed the robustness of fine-tuning against overfitting by studying the two-dimensional generalization error surface of BERT over an extensive range of epochs.

2.3.2 | Evaluation Techniques

Besides the standard general evaluation techniques, the evaluation techniques relevant to our study include both fine-grained evaluation methods and statistical significance testing methods, each contributing different insights into the performance of Portuguese language models.

Fine-Grained Evaluation Fu et al. (2020) have contributed to the evaluation of Chinese Word Segmentation systems by employing a set of attributes for fine-grained evaluation, wherein attributes are used to split the dataset into buckets, which are then evaluated separately. An example on the dataset will be assigned to a bucket if its value for an attribute falls within the range assigned for that bucket.

Statistical Significance Testing Methods The prevalent statistical methods for comparing multiple classifiers across various datasets have been widely discussed in

the literature. Among these methods, Demšar (2006) recommends the utilization of the Friedman test, accompanied by the corresponding post-hoc tests, as the favored technique for these comparisons. However, certain issues arise when applying this method to NLP datasets.

Although the Friedman test does not impose any assumptions regarding the data distribution, it mandates that all test samples be independent. This condition might be violated in text-based data, especially in cases where multiple samples are extracted from a single source. The Friedman test also presents limitations, as it relies on aggregated statistics concerning the metrics, such as averages, without directly considering the measured values themselves. These constraints render the method potentially unsuitable for the analysis of models with high variation, such as deep learning models.

Recognizing these limitations, Dror et al. (2019) introduced an alternative technique named Almost Stochastic Order (ASO). Proven through comprehensive analysis, ASO is particularly adapted to the unique requirements of NLP tasks, providing a more robust framework for significance testing. This innovation mitigates the concerns raised by Dror et al. (2018) regarding the traditional statistical methods, and hence offers a promising direction for more accurate evaluations in the field of NLP.

2.4 | Summary

This chapter provides a comprehensive analysis of developments in Portuguese NLP, with a particular emphasis on the intricacies associated with datasets, benchmarks, and evaluation methodologies for Portuguese language models. We have observed a diverse range of datasets tailored specifically for Portuguese, encompassing applications ranging from semantic similarity to hate speech detection. Noteworthy linguistic phenomena, such as *translationese*, have been identified, highlighting potential pitfalls in benchmark design.

The evolution of Portuguese language models has been traced from their foundational statistical origins to the contemporary paradigm dominated by neural networks. Early innovations like word2vec proved pivotal, establishing a precedent for distributed word representations. There are discernible performance variations when embeddings from one Portuguese variant are applied to content from another, emphasizing the linguistic nuances between dialects of the language.

We have catalogued general-purpose language models available for the Portuguese language. Distinct models such as BERTimbau, PTT5, and Albertina have been elucidated, showcasing their unique contributions and methodologies. The advent of LLMs, exemplified by models like Sabiá-65B, demonstrates emerging capabilities in the realm of in-context learning.

Our exploration also underscores the importance of robust evaluation methodologies in NLP. This includes proper techniques for hyperparameter tuning, seed initialization strategies, and considerations regarding the resilience of transformer-

based models against overfitting during fine-tuning. To gain a holistic understanding of model performance, besides performing a general evaluation, we also highlight the need for fine-grained analysis complemented by rigorous statistical significance testing.

3

Implementation

In this chapter, the architecture and components of the experimental system utilized in this study are delineated, encompassing machine translation, fine-tuning, experiment tracking, and evaluation. A comprehensive flowchart of the system design is depicted in Figure 3.1.

Figure 3.1 delineates the entire system design for our experiments. The process starts when the datasets, both original and translated, and the model checkpoints are transmitted to a GPU Cluster for a three-stage pipeline: hyperparameter optimization (20 trials), random seed selection (40 trials), and fine-tuning (10 trials).

The experiment tracking system encompasses the Weights & Biases Dashboard and a local PostgreSQL database. The former is employed to visualize and monitor model performance, while the latter aids in executing complex queries on the logged data. These tools closely coordinate with the GPU Cluster.

Subsequently, the results are transmitted to an evaluation system, which provides fine-grained evaluation, statistical significance testing, and general evaluation. This facilitates a thorough and robust analysis of the achieved outcomes, which will be reported in the next chapter of this study.

In the following sections, each component of our system will be discussed in detail. This will encompass the hardware and technology employed, as well as the specifics of the implementation process.

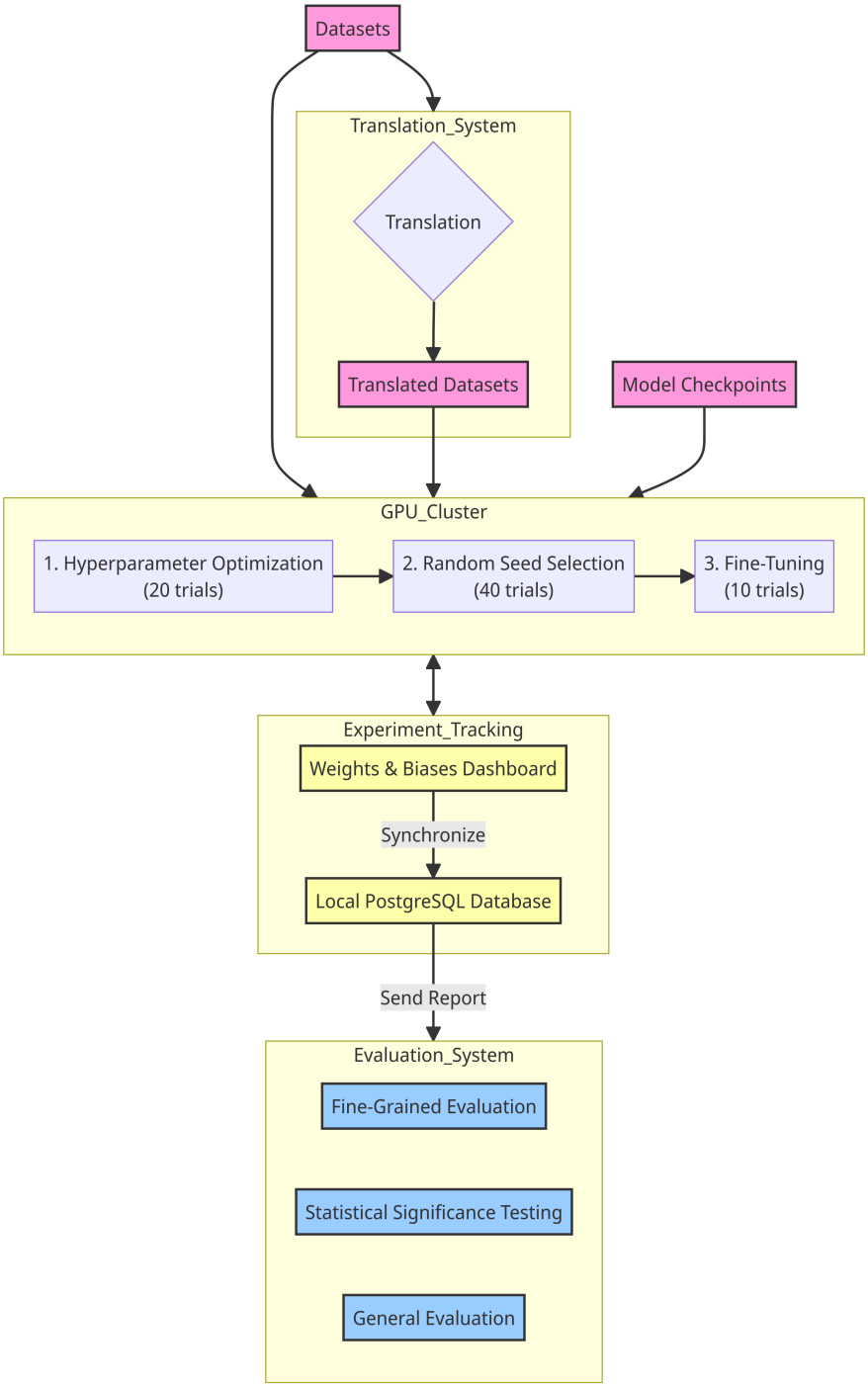


Figure 3.1: Flowchart illustrating the system design for our experiments, encompassing machine translation, training, experiment tracking, and evaluation.

3.1 | Data gathering and pre-processing

This section delineates the data collection and pre-processing steps employed in our study, including details about the datasets and any specific modifications made to them.

The datasets were selected in accordance with the benchmark design criteria detailed in Section 2.1. To ensure the reproducibility of the research and facilitate future developments, all selected datasets have been made publicly accessible as part of the **Na**ural **Po**rtuguese **La**nguage **B**enchmark (**Napolab**) [1].

3.1.1 | Datasets

Table 3.1 presents the selected datasets, along with their respective tasks, task types, and the size of the train, validation, and test splits. The modifications performed on some of the datasets, as detailed below, were necessary to render them suitable for our specific study objectives. Both the ASSIN and ASSIN 2 datasets were employed in their original forms without the necessity for modification.

[1] https://github.com/ruanchaves/napolab

Dataset	Task Name	Task Type	Train	Validation	Test
ASSIN	ASSIN STS	Semantic Similarity	5000	1000	4000
	ASSIN RTE	Textual Entailment	5000	1000	4000
ASSIN 2	ASSIN 2 STS	Semantic Similarity	6500	500	2448
	ASSIN 2 RTE	Textual Entailment	6500	500	2448
HAREM	ReRelEM	Relation Extraction	2226	701	805
PorSimplesSent	PorSimplesSent	Text Simplification	4976	1446	1697
ReLi	ReLi-SA	Sentiment Analysis	7875	1348	3288
HateBR	HateBR	Offensive Language Detection	4480	1120	1400
FaQuaD	FaQuaD-NLI	Question Answering	3128	731	650

Table 3.1: Overview of selected datasets, their corresponding tasks, types, and the number of samples in each data split (Train, Validation, Test).

3.1.1.1 | ReRelEM

docid	sentence1	sentence2	label
hub-56266	Sempre é o iPhone , dirão os que acompanharam as especulações e palpites dos últimos meses . A Apple , empresa de tecnologias que tem nos leitores multimédia iPod e nos computadores [E1]Macintosh[/E1] a sua bandeira, acaba de entrar na área das telecomunicações.	[E2]Apple[/E2] lançou o iPhone , o telemóvel mais esperado dos últimos tempos	produzido_por
hub-28306	-O código malicioso "mais pontual" detectado nos últimos tempos, revela a Panda, foi o Trojan Aifone.A. O nome não engana: é um malware associado ao iPhone da Apple | [E2]iPhone[/E2] da [E1]Apple[/E1].	nan	produtor_de

Table 3.2: Examples from the ReRelEM dataset after modifications. In the first sample(**docid** `hub-56266`), the relationship `produzido_por` suggests the `iPhone` mentioned in **sentence2** is produced by `Apple` from **sentence1**. In the second sample (**docid** `hub-28306`), the `produtor_de` relationship emphasizes that the iPhone is a product of Apple.

The ReRelEM dataset was originally available in XML format with tags annotated over each sample. The annotation format was simplified by repeating the text of each sample multiple times, and annotating only two entities at a time on each resulting sample. The entities on each sample were respectively enclosed by the tags `[E1] [\E1]` and `[E2] [\E2]`, and the relation between them was associated with each sample as its only label. Samples taken from the resulting dataset are shown in Table 3.2. [2]

[2] `https://huggingface.co/datasets/ruanchaves/rerelem`

Training, validation, and test splits were defined for the ReRelEM dataset while ensuring that samples belonging to the same document did not appear in more than one split.

After defining the splits, some instances from the dataset were dropped to ensure consistency: 21 instances with relation labels not included in the training set were dropped from the test set. Additionally, 7 instances from the original dataset that had formatting errors and could not be resolved into post-processed records were also dropped. Since more than 99% of the original instances were retained, it is considered that such changes will not have a significant impact to our experiments.

3.1.1.2 | PorSimplesSent

sentence1	**sentence2**	**label**	**production_id**
Moradores do local afirmam que 11 pessoas foram mortas – e garantem que isso ocorreu antes de a casa ser destruída.	Moradores do local garantem que isso ocorreu antes de a casa ser destruída.	2	3
Moradores do local afirmam que 11 pessoas foram mortas.	Moradores do local afirmam que 11 pessoas foram mortas.	1	3
Propomos colocar números nas peças dos carros que vão para desmanches.	– Propomos mudanças na legislação, com a colocação de números nas peças dos carros que vão para desmanches.	0	4

Table 3.3: Sample entries from the PorSimplesSent dataset. The labels (0, 1, 2) indicate the relative complexity of the sentences in each pair.

The PorSimplesSent dataset was modified by assigning integer labels to each sentence pair, indicating the relationship between the sentences in the pair. These integers (0, 1, 2) represent the relative complexity of the sentences, with 0 indicating that the first sentence is simpler, 1 signifying equivalent complexity, and 2 denoting that the second sentence is simpler. As no standard splits were provided by the original authors, the splits were carefully crafted to prevent sentence pairs from the same document from appearing across multiple splits. Samples taken from the resulting dataset are shown in Table 3.3. [3]

[3] https://huggingface.co/datasets/ruanchaves/porsimplessent

3.1.1.3 | ReLi-SA

unique_review_id	sentence	label
ReLi-Orwell_1984_31	"A história é boa , mas não é um livro que conseguiu prender minha atenção ."	mixed
ReLi-Orwell_1984_32	"Era possível mudar o futuro ?"	neutral
ReLi-Orwell_1984_5	"Sempre aqueles olhos observando a pessoa e a voz a envolvê - la ."	negative

Table 3.4: Examples from the modified ReLi dataset, where token-level sentiment annotations have been transformed to review-level annotations.

The ReLi dataset was originally annotated at the token level, with sentiments attributed to individual tokens. These token-level annotations were transformed into review-level annotations by designating a `positive` label if only positive tokens were identified, a `negative` label if only negative tokens were found, and a `mixed` label if both positive and negative tokens were detected.

As the original dataset had no standard splits, our own splits were defined while carefully avoiding having reviews about a given author appear in more than one split. Samples taken from the resulting dataset are shown in Table 3.4. [4]

3.1.1.4 | HateBR

The annotation format of the original dataset was simplified to boolean annotations based on the features provided. During our experiments, only the binary "offensive language" label was taken into account.

Since the original authors did not define a standard data split, the multi-label data stratification technique implemented in the scikit-multilearn library by Szymański and Kajdanowicz (2017) was utilized to create train-validation-test splits, with the aim of balancing the representation of each hate speech class in the split. Samples taken from the resulting dataset are shown in Table 3.5. [5]

3.1.1.5 | FaQuaD-NLI

The conversion of the question-answering task on the FaQuaD dataset into a textual entailment task (FaQuaD-NLI) was inspired by the methodology employed in the

[4] `https://huggingface.co/datasets/ruanchaves/reli-sa`
[5] `https://huggingface.co/datasets/ruanchaves/hatebr`

instagram_comments	offensive_language
"Fala bandidão"	true
"É uma cara de pau mesmo...	true
"Grandes coisa"	false
"Enquanto existir homem e dinheiro na terra adeus natureza. O ser humano é único animal que destrói o seu próprio hábitat."	false

Table 3.5: Sample comments from the modified HateBR dataset. The original multi-feature annotations were disregarded; instead, only the binary "offensive language" label was used in our experiments.

document_index	document_title	paragraph_index	question	answer	label
2	"CIENCIA DA COMPUTACAO"	0	"Por quem eram produzidas as tabelas logarítmicas?"	"Seu uso original era desenhar linhas na areia com rochas."	0
2	"CIENCIA DA COMPUTACAO"	0	"Qual era o uso original do ábaco?"	"Seu uso original era desenhar linhas na areia com rochas."	1
37	"renovação de matrícula aditivo"	0	"o que a resolução no 401 estabelece?"	"esta resolução entra em vigor na data de sua publicação."	0

Table 3.6: Examples from the modified FaQuaD dataset, transformed into a textual entailment task (FaQuaD-NLI). Each row represents a question-paragraph pair with an associated positive (1) or negative (0) label. A positive label indicates the answer to the question is present in the paragraph, while a negative label denotes the opposite. This adaptation was inspired by the methodology used in the GLUE benchmark for the QNLI dataset.

GLUE benchmark for the QNLI dataset Wang et al. (2018). This transformation simplified the input and output formats for this task.

Instead of indicating the correct answer to a question by enclosing it with tags, we constructed question-paragraph pairs and assigned a `positive` or `negative` label depending on whether the answer to the question was present in the paragraph or not.

As no standard train, validation, and test splits were provided, we defined our own splits carefully to prevent question and answer pairs from the same document from appearing in multiple splits. Samples taken from the resulting dataset are

shown in Table 3.6. [6]

3.2 | Translation Experiments

In this section, we outline experiments designed to investigate the efficacy of fine-tuning our models on the datasets after being translated from Portuguese into other languages. The interest in this comparison emerges from the hypothesis that models pretrained in languages other than Portuguese, such as multilingual models or those specialized in Spanish and Galician (e.g., IXambertv2 and Bertinho, respectively), might display enhanced performance on translated data.

Our experiments begin with the translation of the selected Portuguese datasets into three target languages: English, Spanish, and Galician. The translation process is executed using the NLLB model (No Language Left Behind) with 1.3 billion parameters released by Costa-jussà et al. (2022). All experiments used the default hyperparameters set up by the Easy-Translate library García-Ferrero et al. (2022) as of the release version 2.0, and each experiment was performed on a single dedicated Titan V GPU with 12GB of memory.

Due to the limitations of the NLLB model, the ReRelEM dataset was not included in the translation experiments, and all language models were fine-tuned solely on its Portuguese version. Research into translation systems capable of preserving the entity tags in the ReRelEM dataset is left for future work.

Special attention is paid to the datasets comprised of sentence pairs. In translating these, we take care to translate the pairs as a unified input rather than treating each sentence in isolation. As mentioned under Section 2.1.1.2, this methodological choice helps in minimizing translation artifacts and preserving the lexical overlap between the sentence pairs, thereby maintaining the integrity of the original dataset.

3.3 | GPU Cluster Pipeline

This section explains the method used to fine-tune models on the selected datasets mentioned earlier. The primary objective is to minimize variance in results to ensure fair comparison between different models. The process is divided into three main steps, as detailed in the next subsections.

3.3.1 | Experimental Setup

All experiments were conducted on a single GPU, selected based on the model size. The Albertina and XLM-RoBERTa Large models were fine-tuned on A100 GPUs with 80GB of memory. All other models were fine-tuned using Titan V, Titan X, or Titan Xp GPUs, each with 12GB of memory.

[6] https://huggingface.co/datasets/ruanchaves/faquad-nli

Every experiment is performed by running a modified version of the `run_glue.py` script made available by the Hugging Face library for fine-tuning pre-trained Transformer model on text classification tasks Wolf et al. (2020) [7]. The script was modified to accomodate our experimental steps and to allow for more precise experiment tracking, to be described in the next section.

3.3.2 | Experiment Tracking

We systematically used the Weights & Biases (*wandb*) platform to track all steps of our experiments on the GPU cluster. We conducted a total of 24427 runs on the platform, and our wandb runs are publicly available at `https://wandb.ai/ruan/eplm`.

The data from all wandb runs was synchronized with a local PostgreSQL database to enable customized queries, which are not intrinsically supported by the wandb platform. These queries were executed intermittently between the experimental steps delineated below to configure the parameters for the subsequent step.

3.3.3 | Experimental Steps

In our experimental pipeline, we fine-tune each model on one task and dataset at a time. Experiments involving data augmentation, where fine-tuning would occur on both the original and translated datasets simultaneously, are left for further exploration in future research.

Throughout this subsection, the fine-tuning of the Albertina-PTBR model on the ASSIN 2 RTE dataset is utilized as a representative example. The identical process was applied to all other combinations of models and tasks presented in this study.

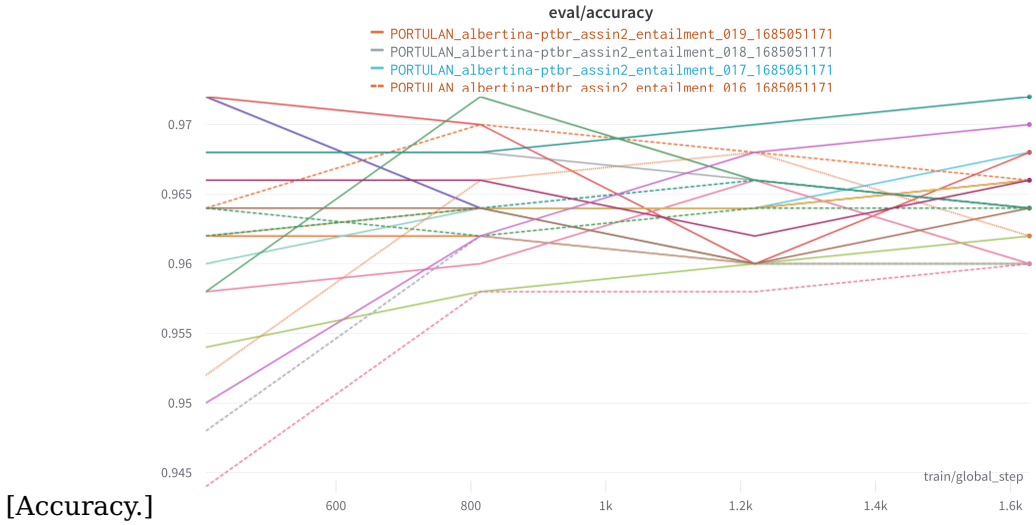
3.3.3.1 | First Step: Hyperparameter Search

The first step in the pipeline is to conduct a quasi-random search for three hyperparameters: learning rate, adam beta1, and warmup steps. We perform 20 trials for each model and task to determine the optimal settings. We pick the hyperparameters of the model that achieves the best performance under the main task metric for the next step.

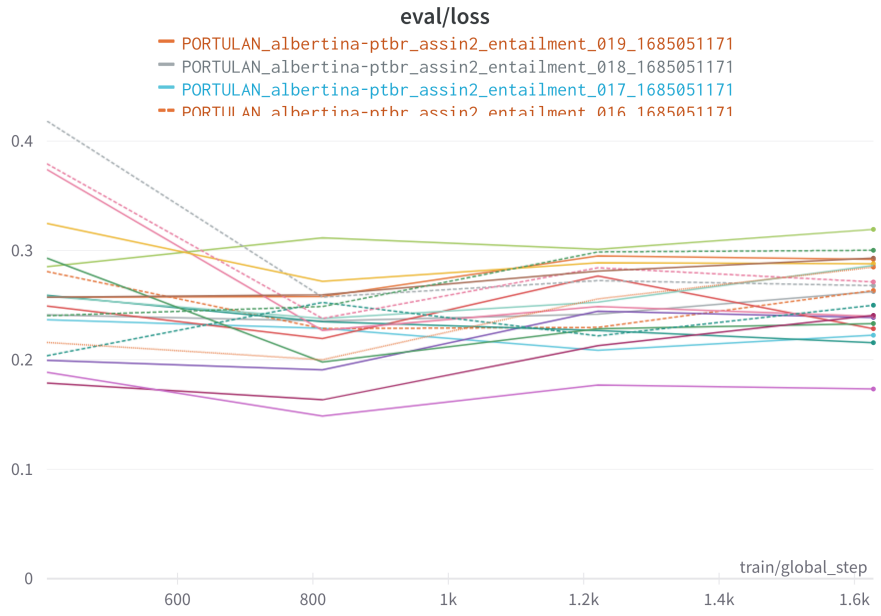
Figure 3.2 illustrates the quasi-random search for the optimal hyperparameters on the validation set of the ASSIN 2 dataset, textual entailment task. Among all trials, the hyperparameters of the model with the highest accuracy were chosen for the next step.

Upon completion of the quasi-random search, the selection of the optimal hyperparameters is undertaken after the wandb platform is synchronized with the local

[7] `https://github.com/huggingface/transformers/tree/686c68f64c9d0181bd54d4d2e2446543c3eca1fa/examples/pytorch/text-classification`



[Accuracy.]



[Loss.]

Figure 3.2: Accuracy and loss plots for the hyperparameter search of Albertina-PTBR on the validation set of the ASSIN 2 dataset, textual entailment task (RTE).

PostgreSQL database. Then, a query is issued to select the best-performing hyperparameters, which are then employed during the generation of the commands for executing the next step.

Search Ranges To ensure the reproducibility of our experiments, all hyperparameters for each trial are documented on the wandb platform, along with the search range within which these hyperparameters were generated by the quasi-random search algorithm.

For the hyperparameter search, we followed the methodology outlined in Section 2.3.1. For any given task and model, we conducted 20 trials at a fixed batch size, employing the quasi-random search implementation recommended by Godbole et al. (2023).

We searched for a total of three hyperparameters: learning rate, Adam beta1, and weight decay. For most of our experiments, the learning rate was searched within the range of 5e-06 to 1e-05, Adam beta1 within the range of 0.5 to 0.999, and weight decay within the range of 0.001 to 0.1. The number of training epochs was always fixed at 4 epochs.

3.3.3.2 | Second Step: Random Seed Selection

Following the identification of the optimal hyperparameters, the fine-tuning procedure is reiterated while altering the random seeds, and maintaining the learning rate, weight decay, and adam beta1 at the best values identified in the preceding step.

In the terminology of Godbole et al. (2023), the first step investigated the *study variance*, while the second step investigates the *trial variance*. By performing a random seed selection during the second step, we adopt the suggested procedure by Godbole et al. (2023) of characterizing the trial-to-trial variance after the study variance has been investigated in the first step.

A total of 40 distinct random seeds for each model and task were tested. These seeds correspond to the first 40 abundant numbers [8]. The choice of abundant numbers was made for ease of reproducibility and to use a consistent number series. It must be noted that this approach is not expected to offer any specific advantage over the more common practice of selecting 40 random seeds from the first 40 natural numbers.

After 2 epochs of fine-tuning with each seed, we discontinue the least promising seeds according to the main evaluation metric for the task at hand, and continue with the 10 best ones to the next step, effectively applying the procedure recommended by Dodge et al. (2020) and described in Section 2.3.1.

3.3.3.3 | Third Step: Final Fine-Tuning

In this phase, the 10 best trials found in the previous step are fine-tuned for 20 epochs. The models are subsequently evaluated on both the validation and test sets of the selected datasets.

Figure 3.2 shows the final fine-tuning on the validation set of the ASSIN 2 dataset, textual entailment task. The results obtained at the end of the 10 trials are then

[8] A number is classified as an abundant number if the sum of its proper divisors exceeds the number. The initial 40 abundant numbers are 12, 18, 20, 24, 30, 36, 40, 42, 48, 54, 56, 60, 66, 70, 72, 78, 80, 84, 88, 90, 96, 100, 102, 104, 108, 112, 114, 120, 126, 132, 138, 140, 144, 150, 156, 160, 162, 168, 174, 176.

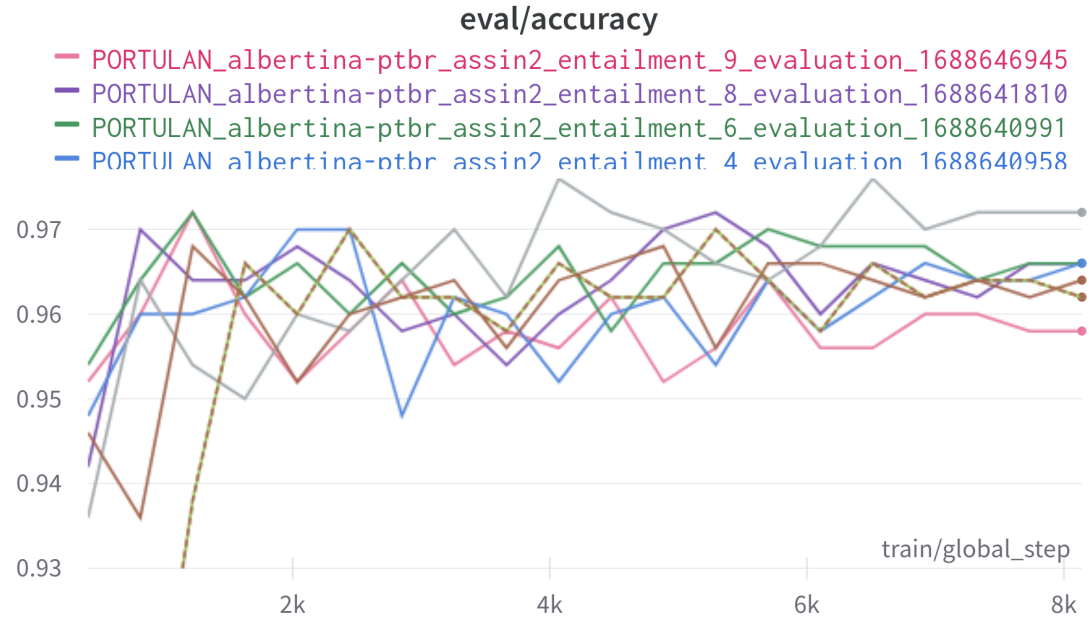
submitted to the component responsible for evaluation in our system.

If this third step reveals an unsatisfactorily high standard deviation between the models on the validation set, we return to the first step after adjusting the search space, and the experimental pipeline is then repeated.

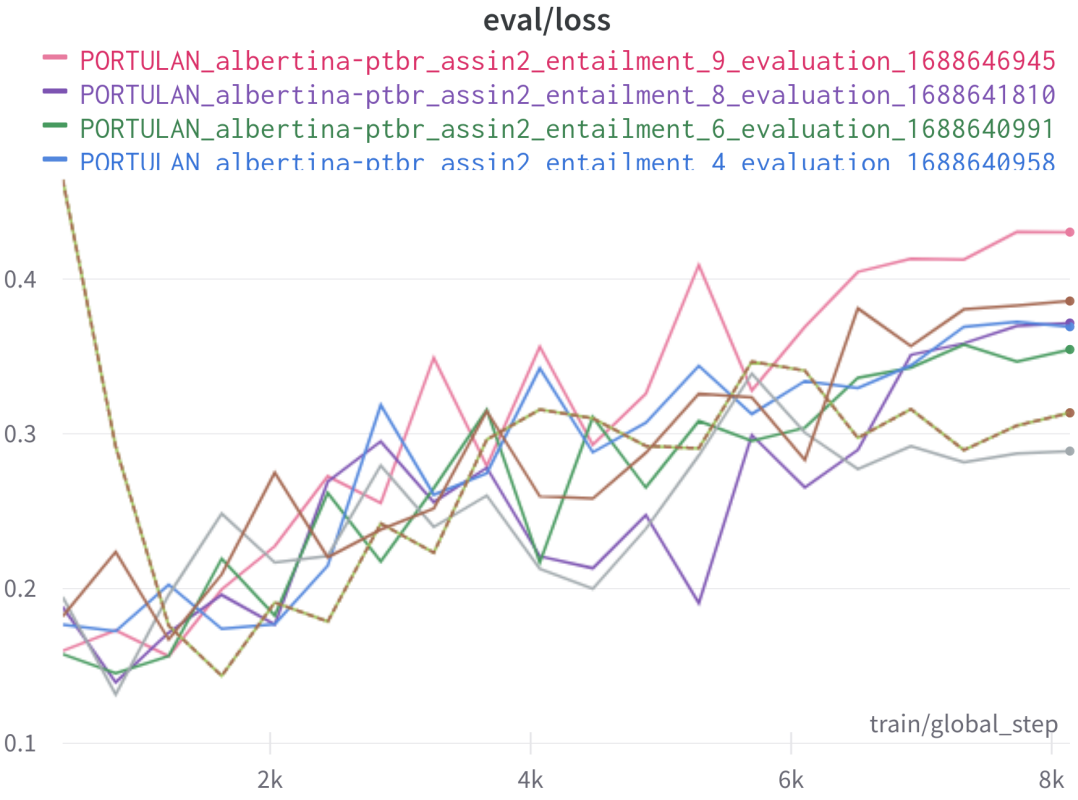
Figure 3.4 highlights this procedure. After finding an unsatisfactorily high standard deviation in the results at the end of the third step, we follow the procedure advocated by Godbole et al. (2023): we choose a single parameter to be adjusted in the next iteration of the pipeline, such as the learning rate. It is important to adjust one hyperparameter at a time to ensure that the results are not affected by multiple changes in the hyperparameter space.

To perform the adjustment, we examine the distribution of the runs in the search space with respect to the chosen hyperparameter. If the best runs are not concentrated in the center of the search space for the chosen hyperparameter, we adjust the hyperparameter accordingly so that the next iteration of the pipeline will be centered in the best region, making it more likely to achieve optimal results.

For instance, we can start our pipeline with large models using a search range for the learning rate of $[1e-05, 5e-05]$. We may notice, however, that this learning rate is too high for large models, which will be reflected in a high standard deviation in the results and a tendency for the best runs to become concentrated in the lower range of the search space for the learning rate. Therefore, for the next iteration of the pipeline, we adjust the learning rate by lowering our search range for the learning rate to $[5e-06, 1e-05]$, while keeping the search range for the other hyperparameters unchanged.



[Accuracy.]



[Loss.]

Figure 3.3: Accuracy and loss plots for the final fine-tuning of Albertina-PTBR on the validation set of the ASSIN 2 dataset, textual entailment task (RTE).

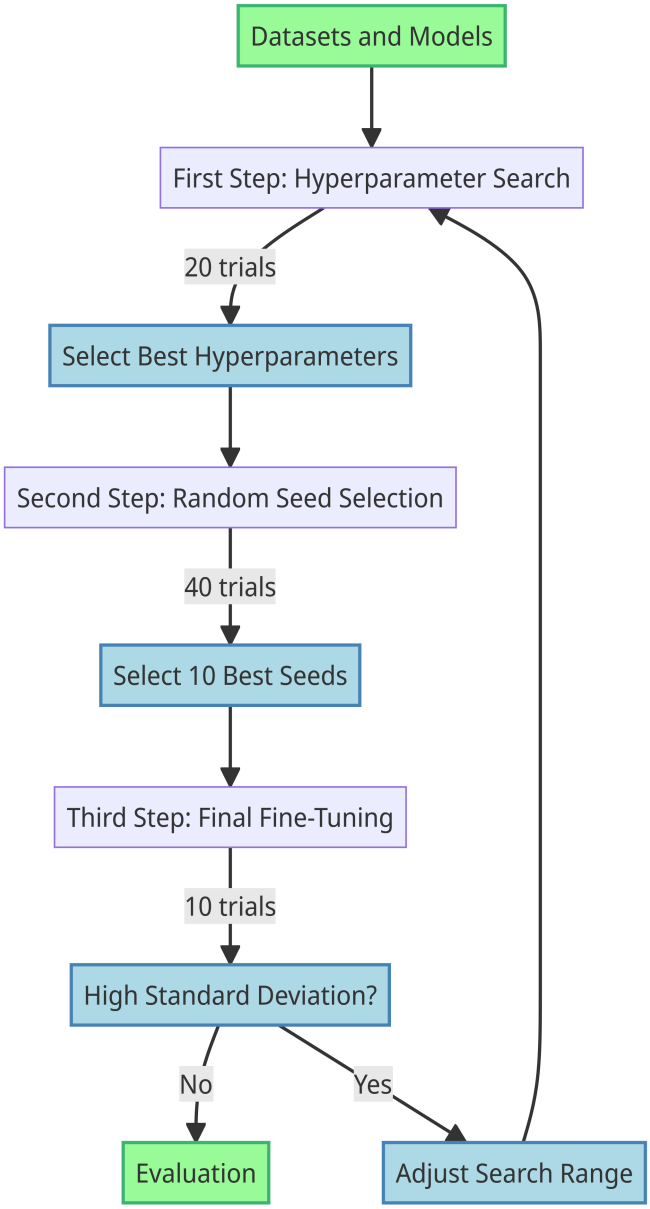


Figure 3.4: Flowchart of the model fine-tuning cycle in our experimental pipeline. Steps highlighted in blue are processed outside the GPU cluster and involve queries to the offline PostgreSQL database, while steps in gray are performed on the GPU Cluster. Hyperparameters are adjusted cyclically until satisfactory performance is achieved.

3.4 | Evaluation

In this section, we delineate the methodology adopted for evaluating the performance of the proposed model in classification tasks. The evaluation is threefold: first, a general evaluation is carried out; second, a fine-grained evaluation is performed to comprehend the nuances of the model's performance in different aspects; and finally, a statistical significance analysis is conducted to ascertain the reliability of the experimental results.

3.4.1 | General Evaluation

For classification tasks, the metrics collected include Macro F1, recall, accuracy, and precision. For regression tasks, such as semantic similarity, the Pearson correlation coefficient and Mean Squared Error are collected.

In the purpose of a general evaluation, all metrics for trials associated with a model and task are averaged. In our evaluation, we also consider the model size, and the trade-off achieved between the results and model size. We also carefully compare the language models with their starting encoders, if they were not pretrained from scratch, to determine if there was an improvement over the original model.

3.4.2 | Fine-Grained Evaluation

We extend the methodology of Fu et al. (2020) mentioned in Section 2.3.1 by developing an attribute-aided fine-grained evaluation method for classification tasks. The proposed method incorporates four essential attributes, as detailed below:

1. **Label Consistency (LC):** Let $C(t_i, L_j)$ denote the count of a specific token, t_i, in the subset of the training set associated with the label L_j. We aim to quantify the label consistency of an example from the test set with respect to a given label, L_j. The *Label Consistency* of a sentence pair in the test set, considering the label L_j, can be defined as:

$$LC_j = \frac{1}{n} \sum_{i=1}^{n} \frac{C(t_i, L_j)}{C(t_i)} \tag{3.1}$$

   where n is the total number of tokens in the sentence pair and $C(t_i)$ is the total count of token t_i in the training set.

2. **Out-of-Vocabulary (OOV) Rate (R_{OOV}):** The OOV rate is determined by calculating the proportion of tokens within a sentence pair in the test set that are not present in the training set. This value is expressed as a percentage.

3. **Word Overlap Rate (R_{WO}):** To ascertain the word overlap rate, we compute the percentage of tokens that are shared between the sentences in a sentence

pair. In other words, the word overlap rate reflects the proportion of common words found in both sentences.

4. **Word Train Frequency (F_{train}):** This attribute represents the average frequency of tokens within a sentence pair in the training set. The frequency is calculated as follows:

$$F_{train} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(t_i)}{f_{max}} \tag{3.2}$$

where n is the number of tokens in the sentence pair, $f(t_i)$ is the frequency of token t_i in the training set, and f_{max} is the maximum frequency of a token in the training set. Tokens not present in the training set are assigned a F_{train} of zero, thus accounting for out-of-vocabulary words.

3.4.3 | Statistical Significance Analysis

In the quest for evaluating the statistical significance of our experimental results, we resort to two complementary approaches. We perform a Friedman test, followed by a Niemanyi post-hoc test. We also perform an analysis of stochastic dominance using a Almost Stochastic Order (ASO) test.

3.4.3.1 | Friedman Test

We conduct a Friedman test exclusively on the results generated by models that were evaluated across all selected datasets. We also restrict our focus to the Portuguese language and exclude experiments with translated data.

For classification tasks, we aggregate the Macro F1 score, recall, precision, and accuracy. For regression tasks, we gather the Pearson correlation and Mean Squared Error scores. Next, for any given model and task, we compute the mean of the test set metrics.

We then subject these averages to a Friedman test followed by a Nemenyi post-hoc test after rejecting the null hypothesis. We consistently set the threshold for the p-value at 0.05 for all tests. We utilize the Friedman test implementation in the pingouin library Vallat (2018), which also facilitates the calculation of the F-value. For the Nemenyi post-hoc test, we employ the implementation in the scikit_posthocs library Terpilowski (2019).

3.4.3.2 | Analysis with Almost Stochastic Order (ASO) Tests

We aggregate the same metrics as mentioned in the previous statistical test. However, instead of computing the average, we provide the raw metrics of each trial as input to the ASO test. We then employ ASO with a confidence level $\alpha = 0.05$ and perform a pair-wise comparison of the performance of all models.

We compare all pairs of models using ASO and adjust all pair-wise comparisons using the Bonferroni correction. We follow the ASO implementation in the deep-significance library Del Barrio et al. (2018); Dror et al. (2019).

3.5 | Summary

This chapter details the implementation of an experimental system designed to evaluate pre-trained language models on Portuguese natural language tasks. The experimental system involved machine translation, fine-tuning, experiment tracking, and evaluation. In our study we present a depiction of the system design, followed by a discussion on data gathering and pre-processing, involving the selection of datasets and their necessary modifications. A thorough methodology for crafting train-validation-test splits, minimizing translation artifacts, and ensuring consistency and balance in class representation was adopted.

Moreover, a GPU Cluster Pipeline was devised for fine-tuning models on selected datasets while minimizing variance in the results. The pipeline involved a three-step process: hyperparameter search, random seed selection, and final fine-tuning. This process was employed to ensure optimal model performance, and if an unsatisfactory standard deviation was observed on the validation set, the training pipeline was repeated with adjusted search space.

The Evaluation section detailed a fine-grained evaluation method developed for classification tasks, incorporating four essential attributes: Label Consistency, Out-of-Vocabulary Rate, Word Overlap Rate, and Word Train Frequency. Additionally, statistical significance analysis was performed using the Friedman test, followed by a Niemanyi post-hoc test and an Almost Stochastic Order (ASO) test.

This chapter therefore provides a comprehensive understanding of the experimental system, datasets, translation experiments, and a detailed experimental pipeline and evaluation methodology for fine-tuning pre-trained language models on Portuguese natural language tasks. Hence, it sets the stage for the analysis of the achieved outcomes in the subsequent chapter.

4

Evaluation

In this chapter, we discuss the results of our experiments. We provide an overall evaluation of the models, conduct a fine-grained evaluation to highlight differences between them, and undertake statistical significance testing to confirm the reliability of our results.

4.1 | General Evaluation

Table 4.1 presents a general summary of the results that will be elaborated upon in the subsequent subsections. It shows DeBERTa v2 (xlarge), a multilingual model discussed in Section 4.4.1.1, which served as the starting checkpoint for the Albertina models. It also displays the multilingual XLM-RoBERTa models, both base and large, the BERTimbau models and their starting checkpoints (the English BERT large for BERTimbau large, and BERT Multilingual for BERTimbau base), and finally, monolingual models pretrained in other languages: Bertinho, which has been pretrained in Galician, and IXAes, which has been pretrained in Spanish.

It's important to underscore that, due to time and resource constraints, not all models were evaluated on every task. Only those models that were evaluated on all tasks were included in the fine-grained evaluation and the statistical significance testing.

Furthermore, it must be emphasized that the fine-tuning process for the ASSIN 2 STS task was not optimal for both the Albertina PT-PT and DeBERTa v2 (xlarge) models. Owing to time constraints, we were unable to continue the hyperparameter optimization cycles within our pipeline to identify the optimal hyperparameters for the fine-tuning of these models on this specific task.

The tables in the next few sections have their rows sorted in descending order according to the metric most relevant for each task. For the semantic similarity tasks, which are regression tasks, they are sorted by the Pearson correlation coefficient. For all other tasks, which are text classification tasks, they are sorted by the Macro

F1 score.

Furthermore, as part of our machine translation experiments, the *language* column in the subsequent subsections represents the language in which each model was fine-tuned and evaluated. For instance, if **spa** is indicated in this column for any given model, it means that the results in that row indicate the performance achieved by this model after fine-tuning on the Spanish-translated training set of the task and evaluation on the corresponding Spanish-translated test set. Similarly, **glg** represents Galician, **eng** stands for English, and **por** signifies Portuguese. Rows with **por** under the *language* column indicate that the model was fine-tuned on the original training set of the task and evaluated on the corresponding original test set.

It's important to note that the tables in the subsequent sections display only the highest-scoring language for each model-task combination. In other words, if Spanish achieves the highest score for a particular model and task, only the results for the Spanish language will be presented in the table for that specific combination. A comprehensive view of the results across all combinations of models, tasks, and languages is available under the Appendix, at Section B.1.

Model	ASSIN RTE	ASSIN STS	ASSIN 2 RTE	ASSIN 2 STS	FaQuaD-NLI	HateBR	PorSimplesSent	ReLi-SA	ReRelEM
Albertina PT-PT	**0.887**	0.874	0.910	0.143	-	-	-	-	-
Albertina PT-BR	0.844	**0.883**	**0.916**	-	-	-	-	-	-
DeBERTa v2 (xlarge)	0.864	0.861	-	0.724	-	-	-	-	-
XLM-RoBERTa (large)	0.874	0.863	0.910	-	-	-	-	-	-
mDeBERTa v3 (base)	0.863	0.855	0.904	**0.847**	0.889	0.911	**0.953**	0.719	0.150
BERTimbau (large)	0.838	0.826	0.897	0.855	**0.900**	**0.919**	0.919	**0.745**	**0.300**
BERT (large)	0.802	0.822	0.892	0.792	0.838	0.838	0.907	0.629	0.113
BERTimbau (base)	0.828	0.844	0.884	0.840	0.897	0.913	0.920	0.713	0.247
BERT multilingual (base)	0.815	0.820	0.877	0.827	0.865	0.871	0.933	0.642	0.183
XLM-RoBERTa (base)	0.822	0.812	0.875	0.817	0.843	0.902	0.929	0.702	0.098
Bertinho	0.786	0.791	0.855	0.802	0.866	0.879	0.900	0.681	0.175
IXAes	0.782	0.817	0.879	0.822	0.860	0.872	0.899	0.637	0.117

Table 4.1: A summary of the general evaluation of language models across various tasks. The table presents the average performance of each model on the test set of each task. The results are displayed according to the most relevant metric for each task, be it the macro F1 Score for classification tasks or the Pearson correlation coefficient for regression tasks. We highlight the highest value on each column. **Blue rows** stand for models that have been pre-trained on multilingual corpora, **green rows** stand for models that have been pre-trained on monolingual corpora in languages other than Portuguese, and rows that are not highlighted with any color stand for models that have been pretrained exclusively on Portuguese data after the starting checkpoint.

4.1.1 | ASSIN (RTE)

In this section, we present the results of all models on the textual entailment task of the ASSIN dataset.

All models underwent fine-tuning using both the Brazilian Portuguese and European Portuguese portions of the dataset, which were treated as a single dataset during fine-tuning.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	Albertina PT-PT	0.878	0.006	0.887	0.004	0.942	0.003
por	XLM-RoBERTa (large)	0.875	0.004	0.874	0.007	0.932	0.004
por	DeBERTa v2 (xlarge)	0.861	0.017	0.864	0.015	0.929	0.009
por	mDeBERTa v3 (base)	0.853	0.010	0.863	0.004	0.927	0.002
por	Albertina PT-BR	0.845	0.117	0.844	0.125	0.932	0.034
por	BERTimbau (large)	0.820	0.017	0.838	0.012	0.920	0.004
por	BERTimbau (base)	0.808	0.013	0.828	0.009	0.917	0.003
por	XLM-RoBERTa (base)	0.809	0.011	0.822	0.006	0.907	0.004
eng	BERT multilingual (base)	0.791	0.011	0.815	0.009	0.904	0.003
eng	BERT (large)	0.781	0.015	0.802	0.012	0.900	0.005
spa	Bertinho	0.772	0.010	0.786	0.010	0.884	0.004
glg	IXAes	0.760	0.011	0.782	0.008	0.886	0.002

Table 4.2: ASSIN entailment results.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	Albertina PT-PT	0.892	0.006	0.899	0.005	0.946	0.003
por	XLM-RoBERTa (large)	0.877	0.005	0.880	0.007	0.932	0.004
por	DeBERTa v2 (xlarge)	0.871	0.013	0.875	0.010	0.931	0.007
por	mDeBERTa v3 (base)	0.858	0.011	0.872	0.006	0.927	0.003
por	Albertina PT-BR	0.848	0.116	0.851	0.122	0.933	0.034
por	BERTimbau (large)	0.825	0.018	0.846	0.011	0.917	0.004
por	XLM-RoBERTa (base)	0.827	0.010	0.843	0.006	0.912	0.005
por	BERTimbau (base)	0.819	0.015	0.838	0.010	0.916	0.005
eng	BERT multilingual (base)	0.813	0.016	0.836	0.013	0.904	0.004
eng	BERT (large)	0.796	0.016	0.816	0.013	0.900	0.006
glg	Bertinho	0.787	0.011	0.807	0.008	0.891	0.003
glg	IXAes	0.783	0.010	0.804	0.006	0.887	0.002

Table 4.3: ASSIN entailment results for PT-PT.

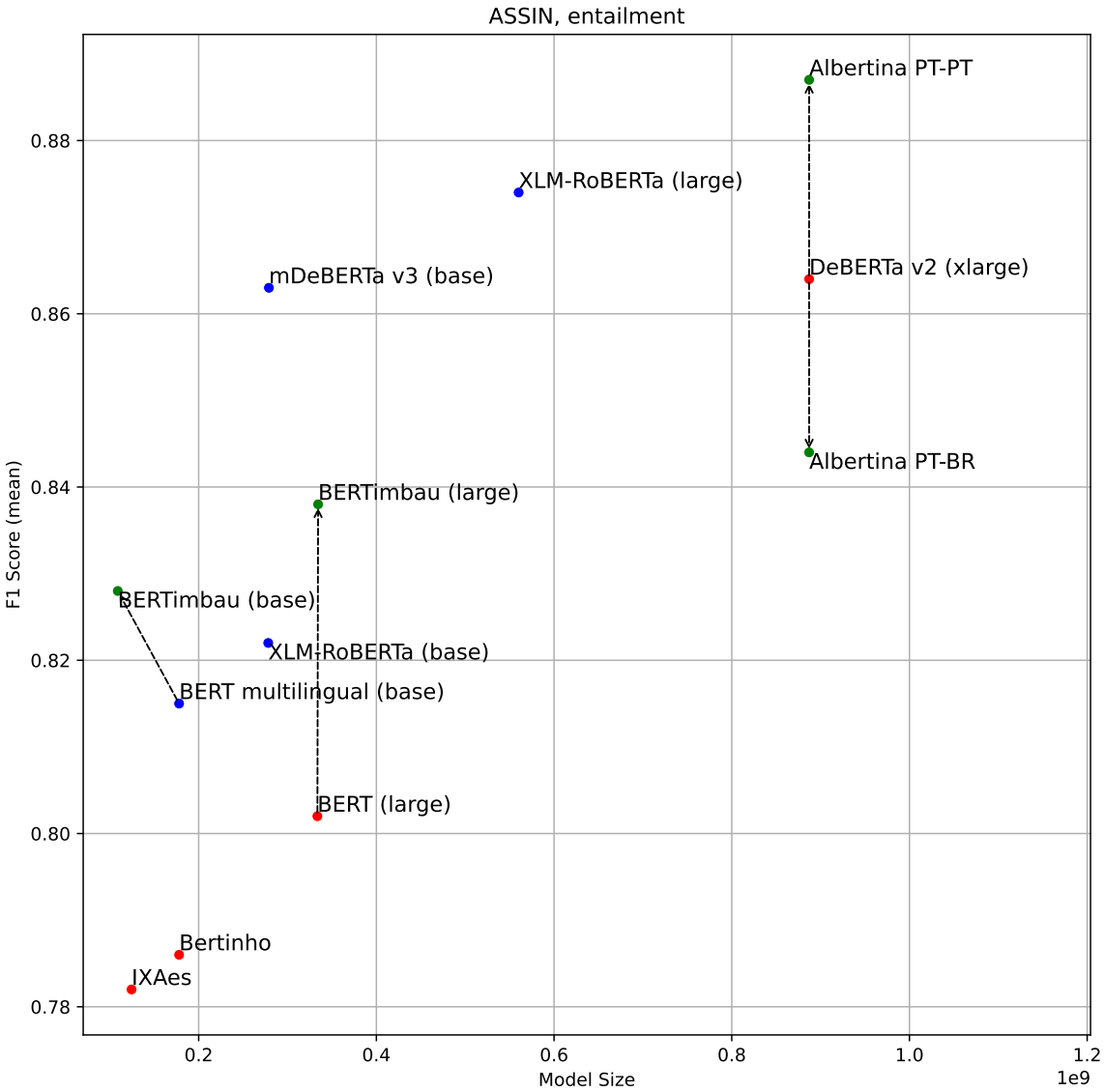


Figure 4.1: ASSIN entailment results, relative to model size (number of parameters).

The results are displayed in Table 4.2, and Figure 4.1 illustrates the mean F-score of each model relative to the number of parameters.

Notably, models pretrained on datasets based on Common Crawl dumps exhibited a significant advantage over similarly-sized models pretrained on the BrWaC corpus. As seen in Figure 4.1, mDeBERTa and XLM-RoBERTa large, which were trained on the CC100 dataset, achieved remarkable scores compared to the BERTimbau Large model pretrained on BrWaC.

Furthermore, an even more prominent observation is that Albertina-PTPT and Albertina-PTBR possess the same number of parameters and were both pretrained starting from the DeBERTa v2 xlarge Transformer encoder. However, Albertina-

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	Albertina PT-PT	0.861	0.009	0.870	0.008	0.938	0.004
por	XLM-RoBERTa (large)	0.872	0.003	0.866	0.009	0.933	0.005
por	mDeBERTa v3 (base)	0.846	0.010	0.851	0.005	0.927	0.004
por	DeBERTa v2 (xlarge)	0.847	0.026	0.848	0.024	0.927	0.011
por	Albertina PT-BR	0.842	0.118	0.835	0.128	0.932	0.034
por	BERTimbau (large)	0.814	0.020	0.827	0.018	0.922	0.006
por	BERTimbau (base)	0.795	0.013	0.814	0.010	0.917	0.004
por	XLM-RoBERTa (base)	0.786	0.016	0.795	0.010	0.902	0.005
eng	BERT multilingual (base)	0.763	0.010	0.787	0.011	0.903	0.004
eng	BERT (large)	0.763	0.017	0.783	0.014	0.901	0.006
spa	Bertinho	0.760	0.015	0.770	0.013	0.886	0.006
glg	IXAes	0.730	0.015	0.753	0.013	0.885	0.003

Table 4.4: ASSIN entailment results for PT-BR.

PTPT was pretrained on a subset of OSCAR, based on Common Crawl dumps, while Albertina-PTBR was pretrained on BrWaC.

Although the models were fine-tuned on both subsets of the ASSIN dataset, we also conducted separate evaluations for each subset. The results are presented in Table 4.3 for the European Portuguese subset of the test set and in Table 4.4 for the Brazilian Portuguese subset.

The results are consistent with those obtained on the full dataset, and Albertina-PTPT achieved a slightly higher performance edge over XLM-RoBERTa large on the European Portuguese subset of the dataset. It is worth noting that even on the Brazilian Portuguese subset, Albertina-PTPT outperformed Albertina-PTBR.

4.1.2 | ASSIN (STS)

The semantic similarity task proposed in the ASSIN dataset includes the same sentence pairs as the entailment task. However, in our evaluation, the performance difference between models pretrained on Common Crawl and BrWaC was not as pronounced as in the entailment task.

Similar to the entailment task, we fine-tuned all models using the full dataset, treating the European Portuguese and Brazilian Portuguese portions as a single split.

Table 4.5 presents a slight advantage in favor of the Albertina-PTBR model, which was trained on BrWaC, while the other models exhibited comparable performance. This trend persisted even when the models were evaluated separately on the European Portuguese test set, as shown in Table 4.6, and the Brazilian Portuguese test

set, as shown in Table 4.7. Notably, the Albertina-PTBR model outperformed the Albertina-PTPT model on the European Portuguese test set, despite the latter being pretrained on European Portuguese data.

language	model	MSE		PCC	
		mean	std	mean	std
por	Albertina PT-BR	0.246	0.002	0.883	0.001
por	Albertina PT-PT	0.303	0.000	0.874	0.000
por	XLM-RoBERTa (large)	0.372	0.015	0.863	0.002
por	DeBERTa v2 (xlarge)	0.323	0.014	0.861	0.001
por	mDeBERTa v3 (base)	0.387	0.014	0.855	0.002
por	BERTimbau (base)	0.337	0.008	0.844	0.003
glg	BERTimbau (large)	0.338	0.013	0.826	0.006
eng	BERT (large)	0.364	0.015	0.822	0.004
por	BERT multilingual (base)	0.342	0.007	0.820	0.003
spa	IXAes	0.392	0.010	0.817	0.001
por	XLM-RoBERTa (base)	0.431	0.027	0.812	0.020
spa	Bertinho	0.442	0.009	0.791	0.001

Table 4.5: ASSIN similarity results.

language	model	MSE		PCC	
		mean	std	mean	std
por	Albertina PT-BR	0.334	0.003	0.905	0.001
por	Albertina PT-PT	0.416	0.000	0.901	0.000
por	DeBERTa v2 (xlarge)	0.432	0.022	0.889	0.001
por	XLM-RoBERTa (large)	0.501	0.022	0.886	0.002
por	mDeBERTa v3 (base)	0.528	0.020	0.877	0.002
por	BERTimbau (base)	0.454	0.010	0.864	0.003
eng	BERT (large)	0.475	0.025	0.847	0.004
glg	BERTimbau (large)	0.441	0.019	0.844	0.008
eng	BERT multilingual (base)	0.465	0.028	0.844	0.003
por	XLM-RoBERTa (base)	0.567	0.033	0.837	0.022
spa	IXAes	0.516	0.015	0.836	0.001
glg	Bertinho	0.544	0.010	0.814	0.004

Table 4.6: ASSIN similarity results for PT-PT.

language	model	MSE		PCC	
		mean	std	mean	std
por	Albertina PT-BR	0.158	0.001	0.891	0.001
por	Albertina PT-PT	0.190	0.000	0.881	0.000
por	XLM-RoBERTa (large)	0.244	0.009	0.870	0.003
por	DeBERTa v2 (xlarge)	0.213	0.006	0.866	0.001
por	mDeBERTa v3 (base)	0.246	0.010	0.866	0.003
por	BERTimbau (base)	0.220	0.007	0.853	0.004
glg	BERTimbau (large)	0.235	0.009	0.835	0.006
por	BERT multilingual (base)	0.245	0.006	0.827	0.004
spa	IXAes	0.268	0.006	0.827	0.002
eng	BERT (large)	0.253	0.006	0.825	0.003
por	XLM-RoBERTa (base)	0.295	0.022	0.817	0.017
spa	Bertinho	0.302	0.004	0.800	0.002

Table 4.7: ASSIN similarity results for PT-BR.

4.1.3 | ASSIN 2 (RTE)

Our results for the entailment task on the ASSIN 2 dataset, presented in Table 4.8, align with previous research findings from Rodrigues et al. (2023) and Souza et al. (2020), as shown in Table 4.9.

While evaluating the ASSIN 2 dataset, Rodrigues et al. (2023) did not include Albertina-PTPT. However, we found that Albertina-PTPT performs at a similar level to Albertina-PTBR, despite ASSIN 2 being a Brazilian Portuguese dataset. Additionally, the multilingual model XLM-RoBERTa Large performed at the same level as the Albertina models.

When using mDeBERTa, we observed results largely equivalent to BERTimbau Large, regardless of whether the dataset was translated into another language before fine-tuning or not. BERT Large also achieved results equivalent to both models when fine-tuned and tested on an English translation of the dataset.

4.1.4 | ASSIN 2 (STS)

Our results achieved in the semantic similarity task of the ASSIN 2 dataset for the BERTimbau models are consistent with the previous research findings of Souza et al. (2020). Specifically, BERTimbau demonstrated a slight advantage over mDeBERTa. As mentioned in Section 4.2.2, BERTimbau Large exhibited stochastic dominance over all evaluated models in this task.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	Albertina PT-BR	0.917	0.005	0.916	0.005	0.917	0.005
por	Albertina PT-PT	0.911	0.004	0.910	0.004	0.911	0.004
por	XLM-RoBERTa (large)	0.910	0.005	0.910	0.005	0.910	0.005
eng	mDeBERTa v3 (base)	0.905	0.004	0.904	0.004	0.905	0.004
por	BERTimbau (large)	0.898	0.005	0.897	0.005	0.898	0.005
eng	BERT (large)	0.892	0.006	0.892	0.006	0.892	0.006
por	BERTimbau (base)	0.885	0.008	0.884	0.008	0.885	0.008
spa	IXAes	0.879	0.003	0.879	0.003	0.879	0.003
spa	BERT multilingual (base)	0.877	0.005	0.877	0.005	0.877	0.005
eng	XLM-RoBERTa (base)	0.876	0.004	0.875	0.004	0.876	0.004
glg	Bertinho	0.856	0.003	0.855	0.003	0.856	0.003

Table 4.8: ASSIN 2 entailment results.

Source	Model	F1 Score	Accuracy
Rodrigues et al. (2023)	Albertina PT-BR	-	0.913
	Albertina PT-PT	-	-
Souza et al. (2020)	BERTimbau (large)	0.900	0.900
	BERTimbau (base)	0.892	0.892

Table 4.9: ASSIN 2 entailment results from previous work.

language	model	MSE		PCC	
		mean	std	mean	std
por	BERTimbau (large)	0.485	0.048	0.855	0.003
por	mDeBERTa v3 (base)	0.616	0.013	0.847	0.002
por	BERTimbau (base)	0.551	0.012	0.840	0.006
eng	BERT multilingual (base)	0.545	0.016	0.827	0.003
spa	IXAes	0.617	0.015	0.822	0.003
por	XLM-RoBERTa (base)	0.696	0.021	0.817	0.004
glg	Bertinho	0.629	0.008	0.802	0.003
eng	BERT (large)	0.604	0.206	0.792	0.105
por	DeBERTa v2 (xlarge)	0.650	0.259	0.724	0.311
por	Albertina PT-PT	1.197	0.289	0.143	0.312

Table 4.10: ASSIN 2 similarity results.

Source	Model	MSE	PCC
Rodrigues et al. (2023)	Albertina PT-BR	-	0.868
	Albertina PT-PT	-	-
Souza et al. (2020)	BERTimbau (large)	0.500	0.852
	BERTimbau (base)	0.580	0.836

Table 4.11: ASSIN 2 similarity results from previous work.

4.1.5 | ReLi-SA

As mentioned in Section 4.2.2, BERTimbau Large demonstrated stochastic dominance over all evaluated models in this task. Surprisingly, automatic translation did not prove beneficial in enhancing the performance of any of the models, even for those that were not pretrained on Portuguese, such as IXAes and Bertinho. Both IXAes and Bertinho achieved their best results when fine-tuned on Portuguese data.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.763	0.014	0.745	0.009	0.818	0.007
por	mDeBERTa v3 (base)	0.734	0.012	0.719	0.011	0.805	0.004
por	BERTimbau (base)	0.714	0.016	0.713	0.010	0.807	0.003
por	XLM-RoBERTa (base)	0.711	0.008	0.702	0.008	0.799	0.004
por	Bertinho	0.672	0.010	0.681	0.008	0.786	0.004
por	BERT multilingual (base)	0.630	0.006	0.642	0.005	0.765	0.005
por	IXAes	0.623	0.045	0.637	0.051	0.771	0.015
eng	BERT (large)	0.620	0.016	0.629	0.012	0.757	0.005

Table 4.12: ReLi-SA results.

4.1.6 | PorSimplesSent

For the PorSimplesSent dataset, we observed that multilingual models ranked above both BERTimbau models. mDeBERTa, BERT Multilingual and XLM-RoBERTa base ranked above both BERTimbau Large and BERTimbau Base. Automatic translation was ineffective in improving the performance of any of the models, including models that were pretrained on other languages.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	mDeBERTa v3 (base)	0.952	0.003	0.953	0.003	0.957	0.003
por	BERT multilingual (base)	0.932	0.005	0.933	0.005	0.938	0.004
por	XLM-RoBERTa (base)	0.928	0.006	0.929	0.006	0.934	0.005
por	BERTimbau (base)	0.918	0.013	0.920	0.013	0.926	0.012
por	BERTimbau (large)	0.918	0.012	0.919	0.012	0.926	0.011
por	BERT (large)	0.906	0.005	0.907	0.006	0.915	0.005
por	Bertinho	0.898	0.003	0.900	0.003	0.908	0.003
por	IXAes	0.899	0.008	0.899	0.008	0.908	0.007

Table 4.13: PorSimplesSent results.

4.1.7 | FaQuaD-NLI

The BERTimbau models demonstrated the highest effectiveness on the FaQuaD-NLI task, with BERTimbau Large achieving the best performance. However, they were closely matched by mDeBERTa. On the other hand, automatic translation proved to be ineffective for Bertinho and IXAes. Interestingly, for mDeBERTa, translating the text into Spanish resulted in improved results.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.907	0.003	0.900	0.005	0.931	0.004
por	BERTimbau (base)	0.898	0.011	0.897	0.011	0.930	0.007
spa	mDeBERTa v3 (base)	0.898	0.010	0.889	0.008	0.923	0.006
por	Bertinho	0.878	0.005	0.866	0.004	0.907	0.003
por	BERT multilingual (base)	0.863	0.009	0.865	0.012	0.909	0.009
por	IXAes	0.870	0.009	0.860	0.010	0.903	0.007
spa	XLM-RoBERTa (base)	0.853	0.009	0.843	0.011	0.891	0.009
spa	BERT (large)	0.840	0.006	0.838	0.006	0.890	0.006

Table 4.14: FaQuaD-NLI results.

4.1.8 | HateBR

In the HateBR dataset, once more, both mDeBERTa and BERTimbau models exhibit comparable performance, with minimal differences. Automatic translation did not yield any noticeable performance improvements for any of the models.

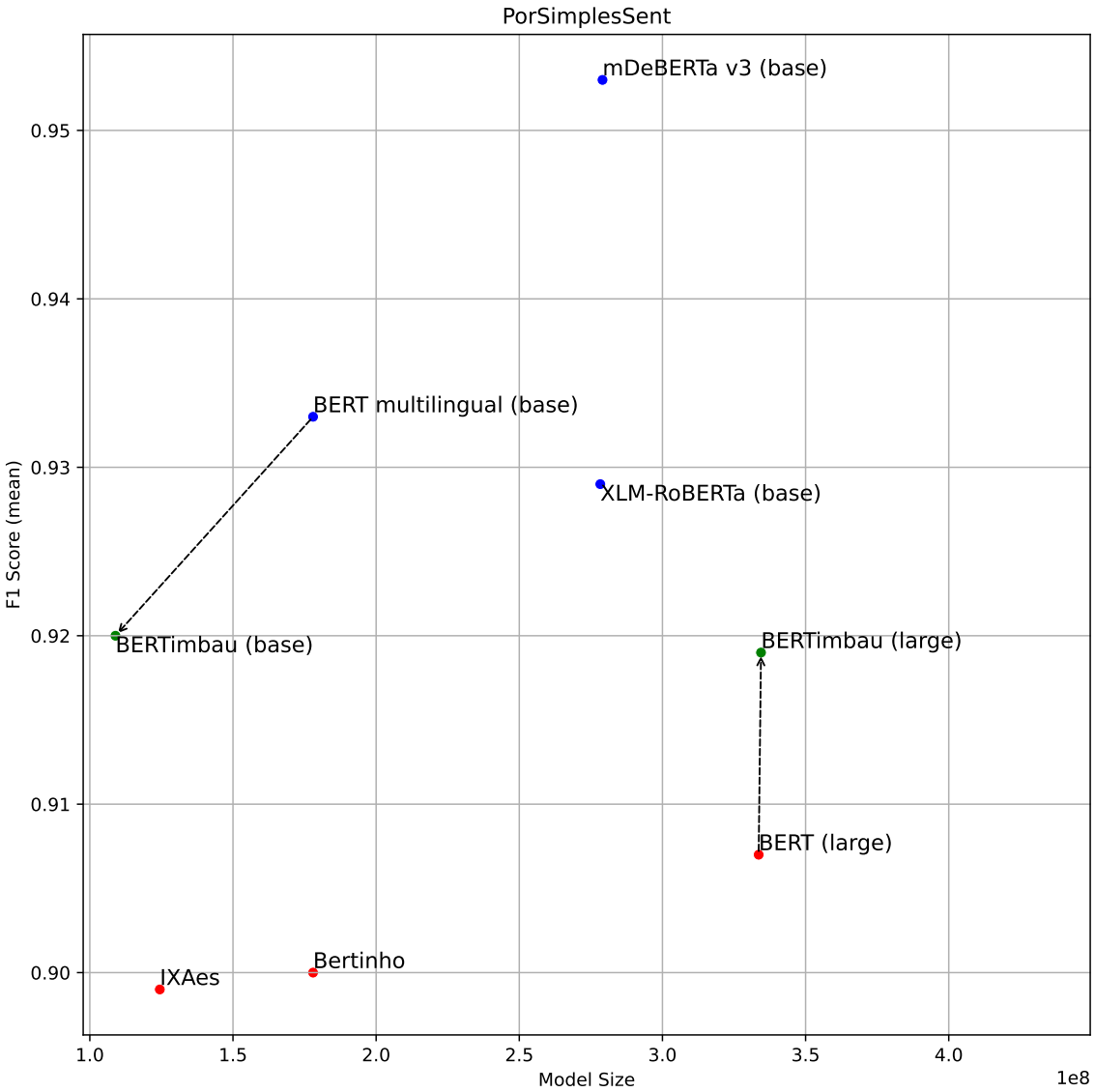


Figure 4.2: PorSimplesSent results, relative to model size (number of parameters).

4.1.9 | ReRelEM

All models exhibited subpar performance on the ReRelEM dataset; however, among them, BERTimbau Large attained the highest results with an F1 score of 0.3 and an accuracy of 0.648. Automatic translation did not prove effective in enhancing the performance of any of the models.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.919	0.005	0.919	0.005	0.919	0.005
por	BERTimbau (base)	0.914	0.003	0.913	0.003	0.914	0.003
por	mDeBERTa v3 (base)	0.911	0.004	0.911	0.004	0.911	0.004
por	XLM-RoBERTa (base)	0.902	0.004	0.902	0.004	0.902	0.004
por	Bertinho	0.879	0.005	0.879	0.005	0.879	0.005
por	IXAes	0.872	0.005	0.872	0.005	0.872	0.005
por	BERT multilingual (base)	0.871	0.007	0.871	0.007	0.871	0.007
por	BERT (large)	0.838	0.054	0.838	0.055	0.838	0.054

Table 4.15: HateBR results.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.316	0.029	0.300	0.032	0.648	0.013
por	BERTimbau (base)	0.260	0.034	0.247	0.033	0.607	0.012
por	BERT multilingual (base)	0.193	0.025	0.183	0.022	0.550	0.026
por	Bertinho	0.189	0.025	0.175	0.024	0.580	0.012
por	mDeBERTa v3 (base)	0.153	0.021	0.150	0.020	0.582	0.014
por	IXAes	0.125	0.009	0.117	0.011	0.539	0.013
por	BERT (large)	0.116	0.007	0.113	0.007	0.535	0.016
por	XLM-RoBERTa (base)	0.112	0.011	0.098	0.012	0.530	0.023

Table 4.16: ReRelEM results.

4.2 | Significance Testing

After fine-tuning ten runs with the best ten random seeds for each model and task at the end of our hyperparameter optimization pipeline, we can conduct significance testing to assess whether there are statistically significant performance differences between the models.

Due to limitations in our computational budget, we will only consider the models fine-tuned on all datasets, which are listed in tables 4.17 and 4.18. Significance testing for the remaining larger models will be left for future work.

We perform two independent statistical experiments: a Friedman test, followed by a post-hoc Nemenyi test, to determine if there is a statistically significant difference between the models; and also an ASO test to determine whether there is a stochastically dominant model among the selected options.

4.2.1 | Friedman test

Following Demšar (2006), we perform a Friedman test to determine whether there are statistically significant differences between the models. We consider Accuracy, Macro F1, Precision and Recall for classification tasks, and Pearson and Mean Squared Error for regression tasks.

For each model, we take the average of the ten random seeds as its value for each metric, and then we perform the Friedman test taking all tasks and selected models into account.

The Friedman test yield a Friedman statistic, $\chi_F^2 = 89.612$, and a p-value $p = 3.648 \times 10^{-17}$, showing that the null hypothesis must be rejected. Therefore, there is a statistically significant difference between the models.

Iman and Davenport (1980) state that χ_F^2 can be undesirably conservative and proposes the F_F statistic. Since $F_F = 27.132$ and the critical value under a significance level of $\alpha = 0.05$ is 0.268, we reject the null hypothesis again because F_F is larger than the critical value.

Given that the Friedman test rejects the null hypothesis, we follow up with a post-hoc Nemenyi test to detect the pairs where there is a statistically significant difference between the models. The results of the Nemenyi test are shown in Table 4.17. We assume a statistically significant difference between two models if their p-value in Table 4.17 is smaller than 0.05.

Two groups of models are clearly distinguished in the table: one group consists of models ostensibly trained on Portuguese data, including mDeBERTa, BERTimbau Base, and BERTimbau Large. The other group comprises models that have been trained on little to no Portuguese data, encompassing all remaining models in the table. Within each of these groups, no statistically significant difference has been detected between the models.

Most importantly, there is no statistically significant difference between the BERTimbau models, which have been exclusively trained on Portuguese data, and a multilingual model trained on Portuguese data along with data from other languages, namely mDeBERTa.

	IXAes	BERT Multilingual	BERTinho	mDeBERTa	BERTimbau (base)	BERTimbau (large)	XLM-RoBERTa (base)
IXAes	**1.000**						
BERT Multilingual	**0.186**	**1.000**					
BERTinho	**0.579**	**0.900**	**1.000**				
mDeBERTa	0.001	0.001	0.001	**1.000**			
BERTimbau (base)	0.001	0.030	0.003	**0.900**	**1.000**		
BERTimbau (large)	0.001	0.001	0.001	**0.900**	**0.817**	**1.000**	
XLM-RoBERTa (base)	**0.511**	**0.900**	**0.900**	0.001	0.004	0.001	**1.000**

Table 4.17: Post-hoc Nemenyi test results. We assume a statistically significant difference between two models if their p-value in this table is smaller than 0.05. It is important to emphasize that all datasets and tasks in our benchmark were included for both the Friedman test and the Post-hoc Nemenyi test results displayed in this table.

4.2.2 | Almost Stochastic Order (ASO)

Almost stochastic dominance ($\epsilon_{min} < \tau$ with $\tau = 0.5$) is indicated in table 4.18. It's important to note that apart from mDeBERTa and BERTimbau Large, none of the models displayed in the table exhibited stochastic dominance over all models in any of the tasks considered in our experiments.

Using Almost Stochastic Order (ASO) we found the score distribution of mDeBERTa to be stochastically dominant over all similarly-sized models on three tasks: ASSIN 2 (RTE), ASSIN (RTE), and PorSimplesSent. BERTimbau Large, under the same settings, was found to be stochastically dominant over all similarly-sized models on three other tasks: ASSIN 2 (STS), HateBR and ReLi-SA.

We compared all pairs of models based on ten random seeds each using ASO with a confidence level of $\alpha = 0.05$, before adjusting for all pair-wise comparisons using the Bonferroni correction. It should be noted that, even if we aim to minimize Type I error (false positives) by considering a smaller value for τ, BERTimbau Large and mDeBERTa still remain stochastically dominant on an equal number of tasks. For $\tau = 0.2$, mDeBERTA is stochastically dominant on PorSimplesSent and ASSIN 2 (RTE), and BERTimbau Large is stochastically dominant on ASSIN 2 (STS) and ReLi-SA.

In summary, ASO revealed stochastic dominance for two models in six out of nine tasks, and both models were stochastically dominant in an equal amount of tasks. Therefore, ASO was not able to provide a clear indication of the dominant model among options pretrained on Portuguese and multilingual datasets, reaching a similar conclusion as the Friedman test and its post-hoc procedures.

Task	Model	IXAes	mBERT	Bertinho	mDeB	BERTim-b	BERTim-l	XLM-R-b
PorSimplesSent	mDeB	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ASSIN 2 (RTE)		0.000	0.000	0.000	0.000	0.000	0.001	0.000
ASSIN (RTE)		0.000	0.030	0.000	0.000	0.136	0.231	0.066
ASSIN 2 (STS)	BERTim-l	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HateBR		0.000	0.000	0.000	0.000	0.002	0.000	0.000
ReLi-SA		0.000	0.000	0.013	0.227	0.183	0.000	0.078

Table 4.18: Almost Stochastic Order (ASO) results. The lower the ϵ_{min} value, the stronger the stochastic dominance of the model under the **Model** column over the model listed as the column name. Stochastic dominance is considered to exist if $\epsilon_{min} < 0.5$, and it is regarded as particularly strong if $\epsilon_{min} < 0.2$.

4.3 | Fine-Grained Evaluation

In this section, we present a fine-grained evaluation of the models. For a given attribute and task, we first categorize the samples in the test set based on the value of that attribute. For example, for the ASSIN 2 dataset, we can categorize samples based on the number of out-of-vocabulary words in each sentence pair. We could assign the interval of zero to one out-of-vocabulary words to the first bucket, one to two out-of-vocabulary words to the second bucket, and so on.

For our study, we've established four buckets for each attribute and task. The specific value intervals for each bucket are detailed in Section B.2.1 under Appendix B. To enhance the interpretability of our results, and following the methodology of Fu et al. (2020), we calculate the Spearman rank correlation coefficient and the standard deviation of each model's performance across all four buckets for each attribute and task. We then compute the global average of the Spearman rank correlation and the standard deviation across all tasks. The aggregate average Spearman rank correlation and standard deviation for each model are displayed in Table 4.20.

A model will exhibit a higher Spearman's rank correlation coefficient for a specific attribute if its performance generally increases as we progress from the first to the last bucket. Conversely, a negative correlation indicates that its performance tends to decline from the first to the last bucket.

The standard deviation of performance across buckets for a particular attribute reflects the model's consistency across various buckets for that attribute. A smaller standard deviation denotes more consistent performance by the model across different attribute values.

4.3.1 | Analysis of Results

Table 4.20 shows an overall advantage for mDeBERTa over the BERTimbau Large and BERTimbau Base models in terms of consistency and variability across most attributes. Notably, the attribute related to the out-of-vocabulary word ratio (R_{OOV}) demonstrates a significant advantage for mDeBERTa.

This advantage, however, is mostly due to the ASSIN and ASSIN 2 datasets. As shown in Table 4.21, if we repeat the same fine-grained evaluation procedures without including the ASSIN and ASSIN 2 datasets, the advantage of mDeBERTa largely disappears, and mDeBERTa and the BERTimbau models become largely equivalent.

The equivalence of the BERTimbau models and mDeBERTa in the fine-grained evaluation becomes further apparent when we visualize the performance of each model under each task and bucket. Figure 4.3 provides a visual overview of the fine-grained evaluation on the ASSIN datasets for the R_{OOV} attribute by showing the fine-grained performance of the most relevant models in our analysis, and Figure 4.4 shows the same for the FaQuAD and HateBR datasets. It can be seen that, although more unstable on the ASSIN datasets, the performance of the models is

largely equivalent on the FaQuAD and HateBR datasets.

In order to better understand this phenomenon, we also examined the distribution of samples across different buckets in the test sets, which helps us understand the prevalence of out-of-vocabulary words in our evaluation. Table 4.19 shows that most test sets in our evaluation predominantly contain samples in the lower buckets, indicating that the words in the test set are largely seen during the training phase. However, the ASSIN 2 dataset is notably skewed, with 96.1% of the samples falling within the first bucket, where the proportion of out-of-vocabulary words in each sentence pair ranges from 0% to 8.3%.

This observation highlights the need for further research on the performance of these models when faced with a higher percentage of out-of-vocabulary words in the test data, which may be more representative of real-world scenarios.

bucket dataset	0	1	2	3
assin	0.658	0.265	0.070	0.007
assin2	0.961	0.030	0.009	0.001
faquad	0.341	0.507	0.133	0.020
hatebr	0.869	0.099	0.030	0.002
porsimplessent	0.559	0.376	0.058	0.006
reli-sa	0.898	0.071	0.028	0.003
rerelem	0.143	0.781	0.068	0.007

Table 4.19: Data distribution of the datasets (in percentages) bucketed by the OOV feature. Columns represent the buckets, while cell values indicate the percentage of data assigned to each bucket.

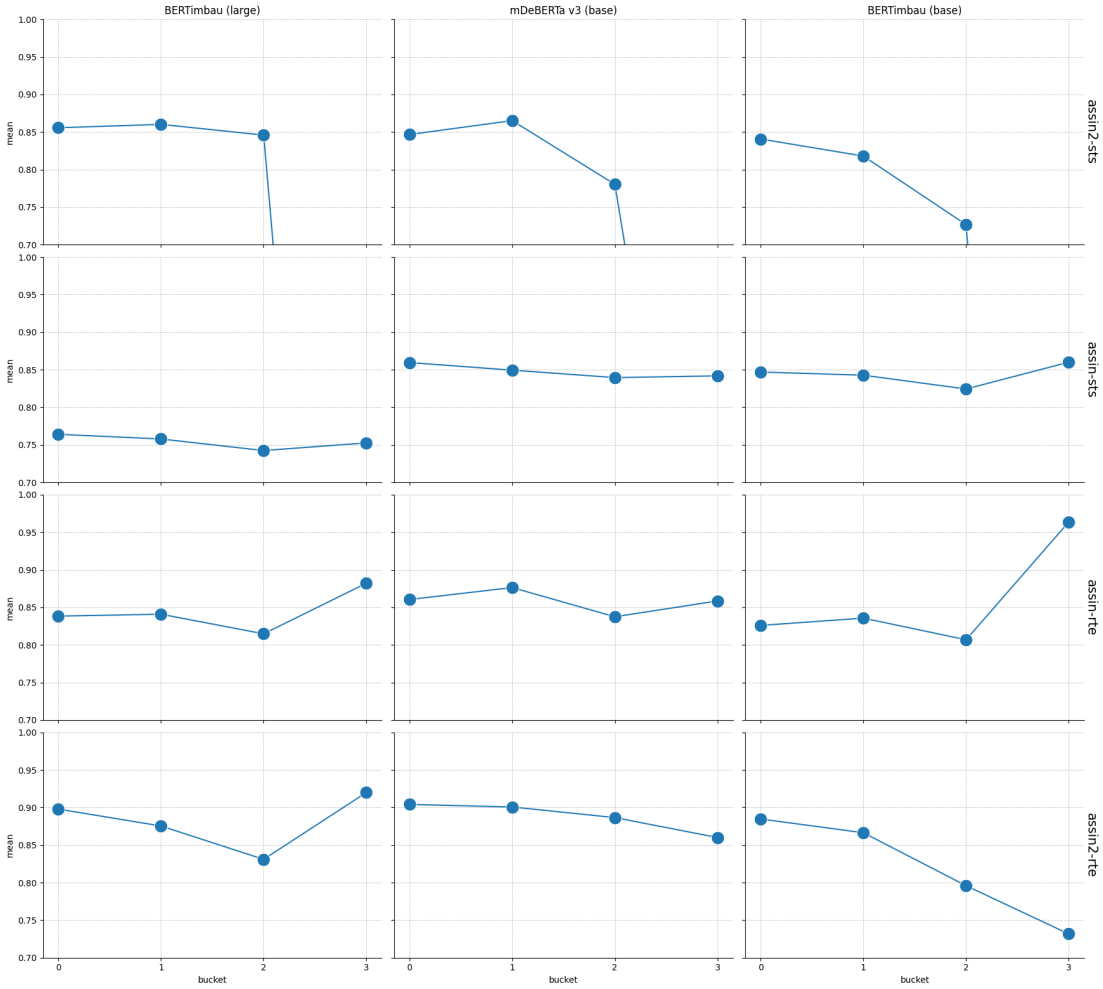


Figure 4.3: Fine-grained performance of models across the ASSIN datasets for the R_{OOV} attribute. The x-axis of each subplot represents the buckets, while the y-axis represents the average performance of each model under the metric most relevant to each task (either F1 Score for classification tasks or Pearson for regression tasks). Values on the y-axis below 0.7 are omitted from the plot. The columns, from left to right, indicate the BERTimbau (large), mDeBERTa, and BERTimbau (base) models. The rows, from top to bottom, represent the ASSIN 2 STS, ASSIN STS, ASSIN RTE, and ASSIN 2 RTE tasks.

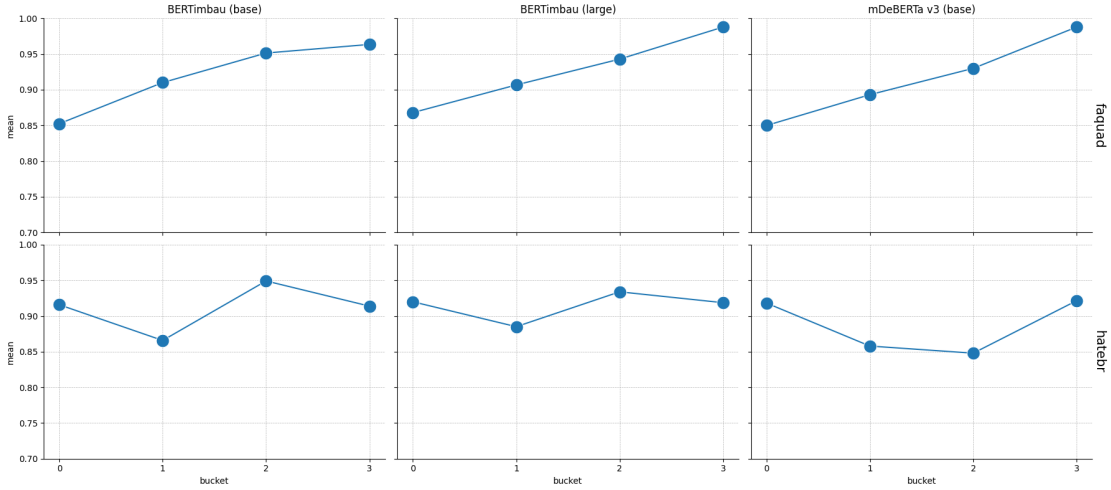


Figure 4.4: Fine-grained performance of models on the FaQuAD and HateBR datasets for the R_{OOV} attribute. The x-axis of each subplot represents the buckets, while the y-axis represents the average performance of each model in terms of the Macro F1 score. The columns, from left to right, indicate the BERTimbau (base), BERTimbau (large), and mDeBERTa models. The rows, from top to bottom, represent the FaQuAD and HateBR tasks.

	spearman					std				
	len	LC	R_{OOV}	R_{WO}	F_{train}	len	LC	R_{OOV}	R_{WO}	F_{train}
model										
mDeBERTa v3 (base)	-0.367	-0.022	-0.467	0.178	0.070	0.069	0.170	0.091	0.050	0.064
BERTimbau (base)	0.056	-0.044	-0.156	0.000	0.292	0.064	0.181	0.167	0.044	0.075
BERTimbau (large)	-0.122	-0.089	-0.200	0.050	0.200	0.065	0.177	0.135	0.055	0.077
IXAes	-0.122	0.000	-0.244	0.067	0.067	0.073	0.191	0.173	0.051	0.087
XLM-RoBERTa (base)	0.199	-0.000	-0.200	0.022	0.194	0.064	0.169	0.135	0.059	0.075
BERT multilingual (base)	0.167	0.000	-0.200	0.055	0.289	0.072	0.188	0.151	0.057	0.085
Bertinho	0.100	-0.044	-0.120	0.089	0.244	0.060	0.187	0.121	0.079	0.081

Table 4.20: Following Fu et al. (2020), we analyze model biases by examining the Spearman's rank correlation coefficient between their performance and the buckets associated with each feature. We also study the standard deviation in performance across buckets for each feature.

	spearman					std				
	len	*LC*	R_{OOV}	R_{WO}	F_{train}	*len*	*LC*	R_{OOV}	R_{WO}	F_{train}
model										
mDeBERTa v3 (base)	-0.60	0.20	-0.20	0.40	0.24	0.101	0.162	0.066	0.056	0.071
BERTimbau (base)	0.00	0.12	0.00	0.20	0.60	0.095	0.177	0.093	0.037	0.089
BERTimbau (large)	-0.04	0.04	-0.16	0.16	0.44	0.095	0.177	0.070	0.060	0.097
IXAes	-0.48	0.16	-0.12	0.08	0.28	0.090	0.196	0.088	0.054	0.111
XLM-RoBERTa (base)	0.24	0.36	-0.12	0.12	0.51	0.092	0.186	0.105	0.051	0.084
BERT multilingual (base)	-0.08	0.12	-0.16	0.12	0.60	0.101	0.193	0.099	0.053	0.105
Bertinho	0.00	0.12	-0.16	0.08	0.48	0.075	0.184	0.090	0.090	0.098

Table 4.21: Fine-grained evaluation results, calculated in the same way as Table 4.20, but excluding the ASSIN and ASSIN 2 datasets.

4.4 | Discussion

This section elucidates the results obtained in the context of existing research and addresses the implications of various facets of our findings.

4.4.1 | Best Practices for Model Evaluation

Drawing from our extensive experiments detailed in this study, we emphasize the importance of specific best practices to ensure an objective and comprehensive evaluation of Portuguese language models:

4.4.1.1 | Incorporating Checkpoints in Evaluation

Models With Unadvertised Portuguese Training During our evaluation, we observed that although some models were pretrained on Portuguese data, this information was not explicitly highlighted in the publications associated with them. For example, as depicted in Table 2.7, DeBERTa v3 was pretrained on a significant corpus of Portuguese data from the CC-News dataset. However, its model card on the Hugging Face Hub identifies it as an English model [1]. Moreover, the experiments reported in its associated paper do not encompass Portuguese tasks He et al. (2021). Thus, its authors did not present it as a model suitable for Portuguese tasks, even though it displays high competitiveness when compared to other alternatives.

Such practices are prevalent in the literature. Consequently, it's vital to be aware of this when choosing models for evaluation. It's necessary to scrutinize the datasets on which models have been trained, investigating the percentage that is in Portuguese. Failing to do so may lead us to inadvertently overlook robust models pretrained on Portuguese data, resulting in an incomplete evaluation.

Starting Checkpoints Given that Portuguese language models are frequently trained from the checkpoint of a pre-existing model, it is crucial to report not only the final results but also those of the starting checkpoint and all intermediate stages.

For example, although Rodrigues et al. (2023) compared Albertina with BERTimbau, they did not compare it with its starting checkpoint, DeBERTa v2, which was already pretrained on a significant corpus of Portuguese data. Our findings indicate that DeBERTa's performance was almost equivalent to that of the Albertina-PTPT and Albertina-PTBR checkpoints. Interestingly, in some cases, the pretraining conducted to produce the models actually resulted in a decline in performance compared to the original DeBERTa v2 checkpoint. This was observed in the entailment tasks of the ASSIN datasets, where Albertina-PTBR was surpassed by DeBERTa v2.

Tokenizers Although not part of our experiments, the performance of multilingual models can be considerably enhanced by merely training a tokenizer for the specific

[1] `https://huggingface.co/microsoft/deberta-v3-base`

language in question and replacing the default tokenizer of the multilingual model with it. Rust et al. (2021) provided compelling evidence supporting this claim. The results of this study should be interpreted with the understanding that the performance of multilingual models can be further optimized by training an appropriate tokenizer.

This aspect was overlooked in a recent study by Larcher et al. (2023), where, despite training a Portuguese tokenizer for an OpenLLama checkpoint, the results for the OpenLLama checkpoint with the replaced tokenizer were not reported. Instead, only the results after pretraining the OpenLLama model with the new tokenizer were reported. This approach does not allow for the isolation of the improvements imparted by the new tokenizer from the effects of continued pretraining of the model.

4.4.1.2 | The Necessity of a Diverse Array of Tasks

Another key consideration is the tasks selected for model evaluation. Our study presents evidence that choosing a diverse array of tasks from native datasets can yield unexpected evaluation results, which may lead to different conclusions than if we had relied solely on translated data or a limited selection of datasets and tasks. For example, Souza et al. (2020) used a set of tasks based on the HAREM and ASSIN datasets, where BERTimbau consistently outperformed BERT Multilingual. However, in our study, we introduced a text simplification task, PorSimplesSent, where the performance of BERT Multilingual was slightly superior to that of the BERTimbau models. Notably, since BERTimbau Base was pretrained from BERT Multilingual, this suggests that the pretraining process resulted in a decline in performance on this specific task.

Some of the datasets employed in this study have been applied to the evaluation of Portuguese language models for the very first time, even though they have been publicly available for a few years. This is particularly true for the ReLi, ReRelEM, and PorSimplesSent datasets. A prevalent trend in academic literature leans towards the utilization of datasets previously employed in other studies, often resorting to automatic translations for their application in Portuguese contexts. We argue that such a convenience-driven approach should be reconsidered in light of the importance of developing robust benchmarks. The initiative to pretrain Portuguese language models should be complemented by rigorous endeavors in the formulation and establishment of comprehensive benchmarks, ensuring the precise evaluation of these models.

4.4.2 | Training Corpora and Requirements for Superior Portuguese Models

Our findings challenge the notion of what is required to train Portuguese models that consistently outperform multilingual alternatives. We observed no statistically

significant difference between mDeBERTA and the BERTimbau models. On a related note, Albertina-PTPT and Albertina-PTBR performed very similarly to the multilingual model XLM-RoBERTa Large on the ASSIN tasks and even quite close to DeBERTa v2 xlarge, which served as the starting point for pretraining the Albertina models. Overall, both Portuguese and multilingual language models exhibited comparable performance.

A detailed evaluation of the models revealed that, under the ASSIN datasets, mDeBERTA exhibited superior out-of-domain generalization performance compared to BERTimbau in some tasks. This draws attention to the differences in training data between these models. mDeBERTA was pretrained on a large multilingual segment of Common Crawl, the Portuguese portion of which was approximately nine times larger than the BrWaC dataset used for pretraining BERTimbau.

Other findings also indicate limitations in the BrWaC dataset when used as the sole source of training data. As noted in the evaluation of the ASSIN RTE task, models trained on Common Crawl outperformed similarly sized models trained exclusively on BrWaC, including both BERTimbau and Albertina-PTBR models. Additional evidence supporting this observation can be found in other research. As mentioned in Section 2.2.1.1, according to Santos et al. (2020), Hartmann et al. (2017) successfully compiled a diverse corpus with greater domain coverage than BrWaC, thereby achieving higher quality static word embeddings than those pretrained on the BrWaC dataset.

Artetxe et al. (2022) conducted a case study on Basque and found that models pretrained on a carefully curated dataset performed worse on downstream tasks than models pretrained on Common Crawl dumps. This underscores the importance of factors such as domain coverage and corpus size, which can outweigh data quality. As illustrated in Table 2.7, Portuguese PLMs have not yet been pretrained on a dataset of comparable size to the Portuguese portion of the Common Crawl dumps used for training multilingual models.

Based on the evidence presented in this work, we argue that pretraining PLMs on large datasets could potentially lead to superior out-of-domain generalization performance, even if it entails a compromise on data quality. We, therefore, advocate for the training of PLMs on corpora similar to those used for recent LLMs in the Portuguese language, such as Sabiá, pretrained on the Portuguese subset of the ClueWeb 2022 dataset, or Cabrita, pretrained on the Portuguese subset of the mC4 dataset Larcher et al. (2023); Pires et al. (2023).

4.4.3 | Effect of Pretraining on Specific Portuguese Variants

Although Rodrigues et al. (2023) argued that pretraining distinct models for different Portuguese variants is empirically justified based on their experiments on datasets automatically translated from the GLUE benchmark, our experiments on the ASSIN datasets yielded different results. European Portuguese models sometimes outper-

formed Brazilian Portuguese models on tasks designed for Brazilian Portuguese and vice versa.

Albertina-PTPT surpassed Albertina-PTBR on Brazilian Portuguese tasks, while Albertina-PTBR outperformed Albertina-PTPT on European Portuguese tasks. This raises the question of whether there is any advantage in pretraining models on data from a specific Portuguese variant.

We believe this warrants a more comprehensive study to determine the extent to which such a difference would be significant, especially including tasks where minor differences between the variants are expected to be most impactful, such as text simplification (using, for instance, the PorSimplesSent dataset) and part of speech tagging (where the MacMorpho dataset could be utilized).

4.5 | Summary

This chapter provides a holistic review of our experimental outcomes, an in-depth evaluation distinguishing model differences, and significance tests to verify the reliability of the findings.

Due to constraints related to time and resources, it was not feasible to evaluate every model on each task. Furthermore, the fine-tuning procedure was suboptimal for the Albertina PT-PT and DeBERTa v2 (xlarge) models in the ASSIN 2 STS task due to time constraints.

Besides a general evaluation of the models, two statistical evaluations were conducted: a Friedman test and a post-hoc Nemenyi test, as well as an Almost Stochastic Order (ASO) test.

The Friedman and post-hoc Nemenyi tests were utilized to discern if a significant performance disparity exists between the models. The outcomes of the Friedman test revealed that no statistically significant difference was observed between equivalent models trained solely on Portuguese data versus those trained on multilingual data.

The Almost Stochastic Order (ASO) test sought to identify if any model stochastically dominated the others. Neither mDeBERTa nor BERTimbau Large exhibited dominance across every task, though they did dominate in individual tasks in equal measures. We also performed a Fine-Grained Evaluation centered on the models' performance against specific attributes.

These findings were critically compared with previous literature. This enabled us to ponder about the best practices for model evaluation: the indispensability of documenting results from every stage of model development, the potential improvement in replacing the default tokenizer in multilingual models with one tailored for the target language, and the criticality of leveraging a diverse set of tasks to gain a robust grasp of a model's natural language understanding capabilities.

Our results also call into question the purported benefits of pretraining models exclusively on data from a particular Portuguese variant. It is advocated that a

more comprehensive study be undertaken, especially focusing on tasks where subtle distinctions between variants could play a pivotal role, such as in text simplification and part of speech tagging.

5

Conclusions

Initially inspired by similar work conducted for the Spanish language, our study undertook a rigorous series of experiments to determine if there is any advantage in pretraining Portuguese language models over their multilingual counterparts, and to what extent such an advantage can be maximized by focusing on specific aspects. In this conclusion, we revisit our research goals, consider how they have been met, and reflect on the limitations of our research. We conclude by outlining future work and providing our final remarks.

5.1 | Achieved Aims and Objectives

5.1.1 | First Objective

> *Constructing a comprehensive evaluation benchmark for Portuguese language models and multilingual models that can consistently assess the performance of each individual model.*

During the course of this research, we constructed the Napolab benchmark. This collection of datasets was rigorously selected through five distinct criteria: Public, Reliable, Natural, Human, and General, as outlined in Section 2.1.2. Special attention was paid to the dataset construction process, and we introduced datasets that were evaluated for the first time on Transformer-based pretrained language models.

Our benchmark was able to uncover new relationships between models. For instance, we discovered that multilingual models often outperformed their Portuguese counterparts on the PorSimplesSent task. This was particularly evident in model pairs such as BERT Multilingual and BERTimbau. Interestingly, this discovery was made for the first time during the evaluation conducted in our study.

5.1.2 | Second Objective

Thoroughly evaluating the performance of Portuguese language models and multilingual models across a diverse array of tasks.

In our evaluation, we considered many models that previously played only a marginal role in language evaluations conducted for the Portuguese language. An example is DeBERTa v3, the base model for the Albertina models. Our evaluation thus advocates for a more comprehensive range of models to be considered. This is especially pertinent for multilingual models, which might perform as competently on Portuguese natural language tasks as models designed specifically for Portuguese.

Our evaluation methodology was also robust. Beyond general assessments, we incorporated fine-grained evaluations and statistical significance testing. This rigorous approach allowed us to reveal insights that might have remained obscured otherwise. One notable finding was the superior out-of-vocabulary performance of mDeBERTa on the ASSIN datasets when compared to the BERTimbau Portuguese language models.

5.1.3 | Third Objective

Assessing the impact of machine translation on the performance of Portuguese language models, multilingual models, and models in languages similar to Portuguese.

We conducted a series of translation experiments to gauge the impact of automatic translation on language models. Contrary to our initial expectations, machine translation was not as effective as anticipated. This was especially true for models pretrained in other languages, such as Bertinho and IXAes. In fact, these models typically showed enhanced performance when fine-tuned using original Portuguese text.

Further analysis revealed that while some datasets, like the ASSIN datasets, were amenable to translation, others, such as PorSimplesSent and the HateBR datasets, showed resistance to current general-purpose translation systems. This shows the importance of native evaluations for Portuguese language models, emphasizing the need to move beyond datasets merely translated from English to Portuguese.

5.2 | Critique and Limitations

We acknowledge that our experiments were, to a certain extent, not as exhaustive as we had hoped due to time and resource constraints. While the current results provide valuable insights, we believe that experiments with larger models could have been expanded to all the datasets on our benchmark. Such an expansion would have allowed for more definitive conclusions, especially concerning the Albertina models.

Arguably, important models and datasets were omitted from our research, again due to time and resource limitations. HAREM, a dataset for named entity recognition, and MacMorpho, a dataset for part of speech tagging, were not part of our evaluation, even though they satisfied all our criteria for inclusion.

Lastly, our experiments didn't allow us to draw definitive conclusions regarding the effect of pretraining Portuguese models on different language variants or identify the specific limitations of cross-variant performance depending on the task. The Portuguese language has seen a gap in studies in this area, and a dedicated research effort is required to provide substantial insights into the optimal approach for pretraining models for specific Portuguese language variants.

5.3 | Future Work

In future work, we plan to strengthen the proposed benchmark and bring even more rigor and depth to our evaluation processes. For instance, in addition to the HAREM and MacMorpho datasets, other datasets can join our benchmark if a reasonable annotation effort is made, such as the B2W-Reviews01 dataset.

By limiting our research scope to PLMs based on masked language modeling, we excluded PTT5, which has demonstrated performance on par with BERTimbau and mDeBERTa. We also didn't consider LLMs, which have been a primary focus in pretraining Portuguese language models. In future work, we should also include LLMs and consider LLM-specific aspects in our evaluation, such as concerns about data contamination and prompt design techniques.

We also intend to explore ways to reduce the time required for our evaluation procedures, given that this was a primary constraint during our research. Future research should aim to draw conclusions as robust as our current ones, but with fewer computational resources and more streamlined pipelines.

Regarding machine translation, one fundamental way the experiments could be improved is by fine-tuning all models simultaneously on all the languages for which the data is available (English, Spanish, Galician, and Portuguese). This data augmentation technique could potentially enhance the performance of multilingual models beyond what we observed in our results. Furthermore, it may also be interesting to train Portuguese tokenizers for multilingual models, as this could improve their performance on Portuguese tasks.

Another concern is to perform research on machine translation models and approaches that can accurately translate the datasets in our benchmark. As we observed in our research, the systems we investigated were incapable of tackling the ReRelEM, HateBR, and PorSimplesSent datasets satisfactorily.

Other potential research directions include designing new splits for existing datasets, especially with a different balance of features, such as the ratio of out-of-vocabulary words in the validation and test sets compared to the training set. We could also delve deeper into existing annotations on selected datasets or endeavor to create

new annotations for them, such as adding named entity annotations to the ASSIN datasets. This would lead to deeper linguistic insights.

5.4 | Final Remarks

Although much attention has been paid to the development of Portuguese language models, very little research has been focused on the elaboration of solid benchmarks for the Portuguese language. This can be partially credited to the current incentive structure in academia, where surpassing the state-of-the-art in previous benchmarks is highly rewarded. This phenomenon is often referred to in the literature as "SOTA-chasing". A side effect of this trend is that not enough attention is given to the merit of the benchmarks themselves and how accurately our benchmarks reflect the reality of language models in real-world production environments.

We hope that our study will encourage further research in this alternative direction and inspire the NLP community to undertake broader and more in-depth evaluations of Portuguese language models. In this way, we can ensure that we are progressing towards building models that are impactful in real-world environments and can transcend the confines of controlled academic experimentation.

A

User Manual

This appendix aims to support the reproducibility of our research. We provide instructions on how to reproduce our experiments, including the fine-tuning and evaluation of the models.

The primary repositories for this study are the Napolab benchmark repository [1], and the Evaluation of Portuguese Language Models repository [2]. While the former will be continuously developed, the latter is intended to be archived, documenting the state of the codebase at the time of study submission. While the repository also contains other code specific to the infrastructure used for the experiments, the scripts mentioned in this User Manual should suffice for reproducing our results in different contexts.

All runs at the Weights & Biases dashboard [3] reference commits in the Evaluation of Portuguese Language Models repository. This repository also contains the code used to generate the figures and tables presented in this thesis.

A.1 | napolab

This section documents the Napolab benchmark repository.

The Napolab repository contains code for the `napolab` library. This library is designed to download and process the datasets used in this study, including their translations into other languages.

To download all datasets, simply run:

```
pip install napolab
python -m napolab
```

This action fetches all datasets from the Hugging Face Hub and saves them as CSVs in your current folder.

[1] https://github.com/ruanchaves/napolab
[2] https://github.com/ruanchaves/evaluation-portuguese-language-models
[3] https://wandb.ai/ruan/eplm

To access the datasets inside a Python environment, in the Hugging Face `datasets` library format, run:

```
from napolab import load_napolab_benchmark
napolab = load_napolab_benchmark(include_translations=True)

benchmark = napolab["datasets"]
translated_benchmark = napolab["translations"]
```

A.2 | evaluation-portuguese-language-models

This section documents the Evaluation of Portuguese Language Models repository.

A.2.1 | Environment Installation

The conda environment for this study is exported to the `environment.yml` file in the Evaluation of Portuguese Language Models repository. The following command creates a conda environment named after the repository and installs all dependencies:

```
conda env create -f environment.yml
```

Please note that `eplm` is a local package, and it can also be installed with the following command at the root of the Evaluation of Portuguese Language Models repository:

```
pip install -e .
```

A.2.2 | Feeding the Local PostgreSQL Database

Running the following command at the root of the Evaluation of Portuguese Language Models repository will fetch all runs from the wandb dashboard and input them into a local PostgreSQL database:

```
python results/fetch_runs.py --project eplm --entity ruan --backend
    postgres
python results/fetch_tables.py --project eplm --username ruan --
   backend postgres
```

Furthermore, this command will log the general evaluation results to the local PostgreSQL database:

```
python results/evaluate.py
```

The PostgreSQL database should be available locally with the username `postgres` and the password `postgres`. The database name is `postgres`.

If you wish to skip these steps, a PostgreSQL dump with all the data is publicly available for download [4]. The dump was created with the following command under PostgreSQL 15.3:

```
docker exec postgres_container pg_dump -U postgres > backup
```

Therefore, it can be restored with the command below:

```
pg_restore -d postgres backup
```

A.2.3 | Running the Experimental Pipeline

The evaluation experiment is divided into three steps. Each step has its designated folder named `experiments/{dataset_name}_{step_name}/config`, containing the settings for each experiment. The remaining experiment folders and files are generated from the settings in the `config` folder. Valid examples are available in the `experiments` folder.

All steps log their results to a wandb project. Hence, before initiating steps 2 and 3, you should run `bash scripts/update_metrics.sh` to update the local metrics database, allowing the wandb data to be processed locally and generate the scripts for the subsequent step.

A.2.3.1 | Step 1: Quasi-Random Search

This step conducts a quasi-random search over learning rate, weight decay, and adam beta1. To generate the `scripts` folder for this step, run:

```
bash scripts/write_experiment_scripts.sh experiments/{
    experiment_name_1}
```

Next, execute every script in the newly generated `experiments/{experiment_name_1}/scripts` folder.

A.2.3.2 | Step 2: Random Seed Search

This step searches for the optimal random seeds. After completing the `scripts` folder for step 1, run:

```
bash scripts/update_metrics.sh
bash scripts/write_seed_scripts.sh experiments/{experiment_name_2}
```

Then, execute every script in the newly generated `experiments/{experiment_name_2}/scripts` folder.

[4] https://archive.org/details/eplm_postgres_202309

A.2.3.3 | Step 3: Final Fine-Tuning

This phase trains the top 10 random seeds for 20 epochs. After finishing the `scripts` folder for step 2, execute:

```
bash scripts/update_metrics.sh
bash scripts/write_evaluation_scripts.sh experiments/{
    experiment_name_3}
```

Then, run every script in the newly generated `experiments/{experiment_name_3}/scripts` folder.

A.2.4 | Evaluation

After executing the entire experimental pipeline, the evaluation scripts can be found in the `results` folder of the `evaluation-portuguese-language-models` repository.

The primary relevant scripts include:

- `results/evaluate_classification.py`: general evaluation for the classification tasks.
- `results/evaluate_classification_fine_grained.py`: fine-grained evaluation for the classification tasks.
- `results/evaluate_assin.py`: general evaluation for the ASSIN datasets.
- `results/evaluate_assin_fine_grained.py`: fine-grained evaluation for the ASSIN datasets.

The files produced by these scripts have been saved on the `results/assin` and `results/classification` folders of the repository.

A.2.5 | Translation Experiments

The scripts required for translating the datasets can be found in the `translate` folder. Instructions for execution are provided in the preface of the script. The sample command below will translate 10 entries from the `assin2` dataset from Portuguese to Spanish:

```
python translate_dataset.py \
--dataset_name 'assin2' --source_lang 'por_Latn' --target_lang '
    spa_Latn' \
--model_name 'facebook/nllb-200-1.3B' --fields 'premise' '
    hypothesis' --translate_together --smoke_test
```

To translate the entire dataset, simply remove the `--smoke_test` argument.

B

Evaluation Results

B.1 | General Evaluation

Earlier in this work, in Section 4.1, we presented the general evaluation results of our experiments, selecting the highest-scoring language (English, Spanish, Galician, or Portuguese) for each model and task. In this appendix, we present the complete results for all the models we evaluated, showcasing their fine-tuned performance across all languages.

As mentioned under Section 3.2, due to limitations in our machine translation model, we did not perform machine translation experiments using the ReRelEM dataset. Therefore, we simply reproduce the same table found in the main body of this study.

B.1.1 | ASSIN (RTE)

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	Albertina PT-PT	0.878	0.006	0.887	0.004	0.942	0.003
por	XLM-RoBERTa (large)	0.875	0.004	0.874	0.007	0.932	0.004
por	DeBERTa v2 (xlarge)	0.861	0.017	0.864	0.015	0.929	0.009
por	mDeBERTa v3 (base)	0.853	0.010	0.863	0.004	0.927	0.002
spa	mDeBERTa v3 (base)	0.836	0.007	0.852	0.005	0.921	0.003
eng	mDeBERTa v3 (base)	0.836	0.010	0.851	0.005	0.923	0.002
glg	mDeBERTa v3 (base)	0.835	0.014	0.851	0.010	0.921	0.005
por	Albertina PT-BR	0.845	0.117	0.844	0.125	0.932	0.034
por	BERTimbau (large)	0.820	0.017	0.838	0.012	0.920	0.004
por	BERTimbau (base)	0.808	0.013	0.828	0.009	0.917	0.003
por	XLM-RoBERTa (base)	0.809	0.011	0.822	0.006	0.907	0.004
glg	BERTimbau (large)	0.789	0.011	0.817	0.008	0.905	0.004
eng	BERT multilingual (base)	0.791	0.011	0.815	0.009	0.904	0.003
spa	BERTimbau (large)	0.789	0.011	0.813	0.009	0.901	0.005
por	BERT multilingual (base)	0.787	0.015	0.812	0.010	0.902	0.003
eng	XLM-RoBERTa (base)	0.798	0.008	0.810	0.006	0.899	0.002
spa	XLM-RoBERTa (base)	0.791	0.012	0.808	0.008	0.898	0.005
spa	BERT multilingual (base)	0.778	0.011	0.806	0.006	0.897	0.003
glg	BERT multilingual (base)	0.775	0.011	0.803	0.009	0.897	0.004
eng	BERT (large)	0.781	0.015	0.802	0.012	0.900	0.005
spa	BERTimbau (base)	0.782	0.007	0.801	0.006	0.896	0.003
glg	BERTimbau (base)	0.772	0.011	0.799	0.006	0.898	0.002
spa	Bertinho	0.772	0.010	0.786	0.010	0.884	0.004
glg	Bertinho	0.762	0.010	0.783	0.008	0.886	0.003
glg	IXAes	0.760	0.011	0.782	0.008	0.886	0.002
eng	BERTimbau (large)	0.765	0.011	0.779	0.008	0.887	0.003
eng	BERTimbau (base)	0.756	0.005	0.772	0.005	0.880	0.002
eng	IXAes	0.767	0.009	0.771	0.007	0.880	0.003
spa	IXAes	0.759	0.150	0.768	0.171	0.890	0.055
por	Bertinho	0.748	0.011	0.765	0.010	0.880	0.004
por	IXAes	0.749	0.006	0.758	0.005	0.877	0.003
eng	Bertinho	0.746	0.006	0.755	0.003	0.869	0.001
glg	XLM-RoBERTa (base)	0.741	0.144	0.751	0.165	0.879	0.051
por	BERT (large)	0.719	0.009	0.734	0.007	0.865	0.006
glg	BERT (large)	0.333	0.000	0.282	0.000	0.735	0.000
spa	BERT (large)	0.333	0.000	0.282	0.000	0.735	0.000

Table B.1: Complete version: ASSIN entailment results.

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	Albertina PT-PT	0.892	0.006	0.899	0.005	0.946	0.003
por	XLM-RoBERTa (large)	0.877	0.005	0.880	0.007	0.932	0.004
por	DeBERTa v2 (xlarge)	0.871	0.013	0.875	0.010	0.931	0.007
por	mDeBERTa v3 (base)	0.858	0.011	0.872	0.006	0.927	0.003
spa	mDeBERTa v3 (base)	0.847	0.009	0.864	0.005	0.923	0.003
glg	mDeBERTa v3 (base)	0.846	0.015	0.862	0.011	0.922	0.006
eng	mDeBERTa v3 (base)	0.845	0.010	0.860	0.007	0.923	0.004
por	Albertina PT-BR	0.848	0.116	0.851	0.122	0.933	0.034
por	BERTimbau (large)	0.825	0.018	0.846	0.011	0.917	0.004
por	XLM-RoBERTa (base)	0.827	0.010	0.843	0.006	0.912	0.005
por	BERTimbau (base)	0.819	0.015	0.838	0.010	0.916	0.005
eng	BERT multilingual (base)	0.813	0.016	0.836	0.013	0.904	0.004
glg	BERTimbau (large)	0.807	0.010	0.833	0.010	0.904	0.005
por	BERT multilingual (base)	0.807	0.015	0.831	0.011	0.902	0.005
spa	XLM-RoBERTa (base)	0.811	0.012	0.830	0.008	0.903	0.004
spa	BERT multilingual (base)	0.800	0.011	0.829	0.007	0.899	0.003
eng	XLM-RoBERTa (base)	0.815	0.008	0.828	0.005	0.902	0.003
spa	BERTimbau (large)	0.802	0.013	0.827	0.010	0.901	0.004
glg	BERT multilingual (base)	0.794	0.010	0.821	0.008	0.898	0.004
spa	BERTimbau (base)	0.799	0.008	0.819	0.008	0.897	0.003
eng	BERT (large)	0.796	0.016	0.816	0.013	0.900	0.006
glg	BERTimbau (base)	0.791	0.009	0.816	0.004	0.896	0.002
glg	Bertinho	0.787	0.011	0.807	0.008	0.891	0.003
glg	IXAes	0.783	0.010	0.804	0.006	0.887	0.002
spa	Bertinho	0.781	0.009	0.799	0.011	0.883	0.005
eng	BERTimbau (large)	0.784	0.017	0.797	0.012	0.886	0.005
eng	IXAes	0.792	0.012	0.795	0.009	0.883	0.004
eng	BERTimbau (base)	0.777	0.007	0.791	0.006	0.878	0.004
por	IXAes	0.775	0.006	0.784	0.005	0.878	0.003
spa	IXAes	0.775	0.156	0.783	0.179	0.888	0.068
por	Bertinho	0.760	0.012	0.780	0.010	0.879	0.005
eng	Bertinho	0.764	0.008	0.772	0.005	0.867	0.002
glg	XLM-RoBERTa (base)	0.753	0.148	0.764	0.173	0.875	0.064
por	BERT (large)	0.739	0.009	0.755	0.009	0.864	0.006
glg	BERT (large)	0.333	0.000	0.273	0.000	0.693	0.000
spa	BERT (large)	0.333	0.000	0.273	0.000	0.693	0.000

Table B.2: Complete version: ASSIN entailment results for PT-PT.

B.1.2 | ASSIN (STS)

language	model	MSE		PCC	
		mean	std	mean	std
por	Albertina PT-BR	0.246	0.002	0.883	0.001
por	Albertina PT-PT	0.303	0.000	0.874	0.000
por	XLM-RoBERTa (large)	0.372	0.015	0.863	0.002
por	DeBERTa v2 (xlarge)	0.323	0.014	0.861	0.001
por	mDeBERTa v3 (base)	0.387	0.014	0.855	0.002
glg	mDeBERTa v3 (base)	0.353	0.006	0.847	0.003
spa	mDeBERTa v3 (base)	0.363	0.009	0.847	0.002
por	BERTimbau (base)	0.337	0.008	0.844	0.003
eng	mDeBERTa v3 (base)	0.357	0.006	0.844	0.002
glg	BERTimbau (large)	0.338	0.013	0.826	0.006
eng	BERT (large)	0.364	0.015	0.822	0.004
por	BERT multilingual (base)	0.342	0.007	0.820	0.003
eng	BERT multilingual (base)	0.362	0.018	0.818	0.002
spa	IXAes	0.392	0.010	0.817	0.001
por	XLM-RoBERTa (base)	0.431	0.027	0.812	0.020
spa	BERT multilingual (base)	0.372	0.023	0.805	0.006
spa	XLM-RoBERTa (base)	0.451	0.022	0.805	0.005
eng	XLM-RoBERTa (base)	0.446	0.012	0.804	0.003
glg	BERTimbau (base)	0.389	0.007	0.803	0.007
glg	XLM-RoBERTa (base)	0.456	0.015	0.796	0.009
glg	BERT multilingual (base)	0.394	0.024	0.793	0.007
spa	BERTimbau (base)	0.411	0.017	0.792	0.009
spa	Bertinho	0.442	0.009	0.791	0.001
glg	Bertinho	0.422	0.008	0.789	0.004
por	Bertinho	0.437	0.007	0.778	0.003
glg	IXAes	0.447	0.008	0.774	0.002
spa	BERTimbau (large)	0.433	0.204	0.764	0.163
por	IXAes	0.479	0.014	0.764	0.003
por	BERTimbau (large)	0.374	0.225	0.760	0.293
eng	IXAes	0.489	0.011	0.754	0.003
por	BERT (large)	0.442	0.020	0.752	0.005
eng	Bertinho	0.490	0.012	0.747	0.003
glg	BERT (large)	0.445	0.006	0.742	0.004
spa	BERT (large)	0.496	0.165	0.688	0.189

Table B.3: Complete version: ASSIN similarity results.

language	model	MSE		PCC	
		mean	std	mean	std
por	Albertina PT-BR	0.334	0.003	0.905	0.001
por	Albertina PT-PT	0.416	0.000	0.901	0.000
por	DeBERTa v2 (xlarge)	0.432	0.022	0.889	0.001
por	XLM-RoBERTa (large)	0.501	0.022	0.886	0.002
por	mDeBERTa v3 (base)	0.528	0.020	0.877	0.002
eng	mDeBERTa v3 (base)	0.473	0.008	0.872	0.002
glg	mDeBERTa v3 (base)	0.476	0.010	0.871	0.003
spa	mDeBERTa v3 (base)	0.495	0.014	0.871	0.003
por	BERTimbau (base)	0.454	0.010	0.864	0.003
eng	BERT (large)	0.475	0.025	0.847	0.004
glg	BERTimbau (large)	0.441	0.019	0.844	0.008
eng	BERT multilingual (base)	0.465	0.028	0.844	0.003
por	BERT multilingual (base)	0.440	0.013	0.842	0.003
por	XLM-RoBERTa (base)	0.567	0.033	0.837	0.022
spa	IXAes	0.516	0.015	0.836	0.001
eng	XLM-RoBERTa (base)	0.578	0.017	0.830	0.002
spa	XLM-RoBERTa (base)	0.599	0.031	0.828	0.006
spa	BERT multilingual (base)	0.481	0.034	0.827	0.006
glg	BERTimbau (base)	0.503	0.011	0.822	0.008
glg	BERT multilingual (base)	0.502	0.035	0.817	0.008
glg	XLM-RoBERTa (base)	0.599	0.020	0.817	0.009
glg	Bertinho	0.544	0.010	0.814	0.004
spa	Bertinho	0.582	0.014	0.812	0.001
spa	BERTimbau (base)	0.537	0.022	0.811	0.009
por	Bertinho	0.566	0.011	0.802	0.004
glg	IXAes	0.565	0.012	0.795	0.003
por	IXAes	0.609	0.020	0.787	0.003
spa	BERTimbau (large)	0.560	0.236	0.781	0.174
eng	IXAes	0.597	0.020	0.781	0.004
por	BERTimbau (large)	0.493	0.257	0.777	0.296
por	BERT (large)	0.549	0.028	0.774	0.006
eng	Bertinho	0.608	0.017	0.773	0.003
glg	BERT (large)	0.542	0.007	0.765	0.004
spa	BERT (large)	0.611	0.194	0.711	0.199

Table B.4: Complete version: ASSIN similarity results for PT-PT.

language	model	MSE		PCC	
		mean	std	mean	std
por	Albertina PT-BR	0.158	0.001	0.891	0.001
por	Albertina PT-PT	0.190	0.000	0.881	0.000
por	XLM-RoBERTa (large)	0.244	0.009	0.870	0.003
por	DeBERTa v2 (xlarge)	0.213	0.006	0.866	0.001
por	mDeBERTa v3 (base)	0.246	0.010	0.866	0.003
spa	mDeBERTa v3 (base)	0.232	0.006	0.859	0.002
glg	mDeBERTa v3 (base)	0.230	0.004	0.855	0.003
por	BERTimbau (base)	0.220	0.007	0.853	0.004
eng	mDeBERTa v3 (base)	0.240	0.005	0.849	0.003
glg	BERTimbau (large)	0.235	0.009	0.835	0.006
por	BERT multilingual (base)	0.245	0.006	0.827	0.004
spa	IXAes	0.268	0.006	0.827	0.002
eng	BERT (large)	0.253	0.006	0.825	0.003
eng	BERT multilingual (base)	0.258	0.008	0.819	0.002
por	XLM-RoBERTa (base)	0.295	0.022	0.817	0.017
spa	XLM-RoBERTa (base)	0.304	0.012	0.814	0.003
spa	BERT multilingual (base)	0.264	0.013	0.813	0.007
glg	BERTimbau (base)	0.274	0.005	0.810	0.006
eng	XLM-RoBERTa (base)	0.313	0.009	0.807	0.004
spa	BERTimbau (base)	0.286	0.012	0.803	0.009
glg	XLM-RoBERTa (base)	0.314	0.011	0.801	0.008
spa	Bertinho	0.302	0.004	0.800	0.002
glg	BERT multilingual (base)	0.287	0.013	0.796	0.007
glg	Bertinho	0.299	0.007	0.791	0.005
por	Bertinho	0.307	0.004	0.783	0.003
glg	IXAes	0.328	0.005	0.778	0.003
spa	BERTimbau (large)	0.306	0.171	0.776	0.155
por	BERTimbau (large)	0.255	0.193	0.770	0.301
por	IXAes	0.349	0.010	0.770	0.004
por	BERT (large)	0.335	0.012	0.755	0.005
eng	IXAes	0.380	0.008	0.752	0.004
eng	Bertinho	0.372	0.008	0.745	0.004
glg	BERT (large)	0.348	0.006	0.742	0.005
spa	BERT (large)	0.381	0.136	0.690	0.186

Table B.5: Complete version: ASSIN similarity results for PT-BR.

B.1.3 | ASSIN 2 (RTE)

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	Albertina PT-BR	0.917	0.005	0.916	0.005	0.917	0.005
por	Albertina PT-PT	0.911	0.004	0.910	0.004	0.911	0.004
por	XLM-RoBERTa (large)	0.910	0.005	0.910	0.005	0.910	0.005
eng	mDeBERTa v3 (base)	0.905	0.004	0.904	0.004	0.905	0.004
por	mDeBERTa v3 (base)	0.904	0.003	0.904	0.003	0.904	0.003
spa	mDeBERTa v3 (base)	0.900	0.004	0.900	0.004	0.900	0.004
por	BERTimbau (large)	0.898	0.005	0.897	0.005	0.898	0.005
glg	mDeBERTa v3 (base)	0.897	0.003	0.897	0.003	0.897	0.003
eng	BERT (large)	0.892	0.006	0.892	0.006	0.892	0.006
por	BERTimbau (base)	0.885	0.008	0.884	0.008	0.885	0.008
spa	IXAes	0.879	0.003	0.879	0.003	0.879	0.003
spa	BERT multilingual (base)	0.877	0.005	0.877	0.005	0.877	0.005
spa	BERTimbau (large)	0.877	0.004	0.877	0.004	0.877	0.004
eng	XLM-RoBERTa (base)	0.876	0.004	0.875	0.004	0.876	0.004
spa	XLM-RoBERTa (base)	0.872	0.006	0.872	0.006	0.872	0.006
por	BERT multilingual (base)	0.870	0.004	0.869	0.004	0.870	0.004
glg	BERT multilingual (base)	0.865	0.006	0.864	0.006	0.865	0.006
spa	BERTimbau (base)	0.864	0.005	0.863	0.005	0.864	0.005
glg	BERTimbau (large)	0.863	0.008	0.862	0.008	0.863	0.008
glg	Bertinho	0.856	0.003	0.855	0.003	0.856	0.003
spa	Bertinho	0.854	0.006	0.853	0.006	0.854	0.006
glg	XLM-RoBERTa (base)	0.853	0.014	0.852	0.014	0.853	0.014
eng	BERTimbau (base)	0.849	0.005	0.848	0.006	0.849	0.005
por	Bertinho	0.849	0.005	0.848	0.005	0.849	0.005
glg	BERTimbau (base)	0.846	0.008	0.846	0.008	0.846	0.008
eng	IXAes	0.839	0.006	0.838	0.006	0.839	0.006
por	IXAes	0.839	0.005	0.838	0.005	0.839	0.005
por	BERT (large)	0.838	0.008	0.838	0.008	0.838	0.008
glg	IXAes	0.833	0.005	0.832	0.005	0.833	0.005
eng	Bertinho	0.830	0.005	0.828	0.005	0.830	0.005
eng	BERTimbau (large)	0.705	0.180	0.638	0.265	0.705	0.180
por	XLM-RoBERTa (base)	0.687	0.172	0.618	0.254	0.687	0.172

Table B.6: Complete version: ASSIN 2 entailment results.

B.1.4 | ASSIN 2 (STS)

language	model	MSE		PCC	
		mean	std	mean	std
por	BERTimbau (large)	0.485	0.048	0.855	0.003
por	mDeBERTa v3 (base)	0.616	0.013	0.847	0.002
eng	mDeBERTa v3 (base)	0.597	0.019	0.841	0.002
spa	mDeBERTa v3 (base)	0.617	0.015	0.841	0.003
por	BERTimbau (base)	0.551	0.012	0.840	0.006
glg	mDeBERTa v3 (base)	0.612	0.014	0.840	0.004
eng	BERT multilingual (base)	0.545	0.016	0.827	0.003
spa	BERTimbau (large)	0.523	0.044	0.826	0.009
glg	BERTimbau (large)	0.521	0.032	0.825	0.008
spa	BERT multilingual (base)	0.542	0.020	0.823	0.005
spa	IXAes	0.617	0.015	0.822	0.003
por	BERT multilingual (base)	0.555	0.014	0.817	0.006
por	XLM-RoBERTa (base)	0.696	0.021	0.817	0.004
eng	XLM-RoBERTa (base)	0.644	0.025	0.815	0.006
spa	BERTimbau (base)	0.580	0.011	0.812	0.006
glg	BERT multilingual (base)	0.555	0.028	0.810	0.008
spa	XLM-RoBERTa (base)	0.672	0.027	0.807	0.007
eng	BERTimbau (large)	0.550	0.016	0.803	0.007
glg	XLM-RoBERTa (base)	0.702	0.027	0.803	0.007
glg	Bertinho	0.629	0.008	0.802	0.003
spa	Bertinho	0.644	0.017	0.797	0.004
glg	BERTimbau (base)	0.667	0.021	0.794	0.005
eng	BERT (large)	0.604	0.206	0.792	0.105
por	Bertinho	0.699	0.016	0.792	0.002
eng	BERTimbau (base)	0.627	0.011	0.787	0.005
por	BERT (large)	0.661	0.042	0.770	0.007
eng	IXAes	0.757	0.025	0.770	0.004
eng	Bertinho	0.693	0.010	0.768	0.003
por	IXAes	0.788	0.020	0.760	0.006
glg	IXAes	0.801	0.046	0.757	0.007
por	DeBERTa v2 (xlarge)	0.650	0.259	0.724	0.311
por	Albertina PT-PT	1.197	0.289	0.143	0.312

Table B.7: Complete version: ASSIN 2 similarity results.

B.1.5 | ReLi-SA

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.763	0.014	0.745	0.009	0.818	0.007
por	mDeBERTa v3 (base)	0.734	0.012	0.719	0.011	0.805	0.004
por	BERTimbau (base)	0.714	0.016	0.713	0.010	0.807	0.003
por	XLM-RoBERTa (base)	0.711	0.008	0.702	0.008	0.799	0.004
por	Bertinho	0.672	0.010	0.681	0.008	0.786	0.004
glg	BERTimbau (large)	0.642	0.009	0.643	0.009	0.759	0.005
por	BERT multilingual (base)	0.630	0.006	0.642	0.005	0.765	0.005
por	IXAes	0.623	0.045	0.637	0.051	0.771	0.015
spa	IXAes	0.639	0.009	0.636	0.005	0.754	0.003
spa	mDeBERTa v3 (base)	0.639	0.015	0.634	0.008	0.746	0.003
eng	BERT (large)	0.620	0.016	0.629	0.012	0.757	0.005
eng	mDeBERTa v3 (base)	0.625	0.012	0.629	0.009	0.750	0.004
glg	mDeBERTa v3 (base)	0.631	0.012	0.628	0.011	0.745	0.007
glg	XLM-RoBERTa (base)	0.630	0.011	0.626	0.009	0.741	0.006
eng	XLM-RoBERTa (base)	0.628	0.012	0.622	0.009	0.745	0.005
glg	Bertinho	0.612	0.005	0.618	0.007	0.743	0.005
spa	BERTimbau (large)	0.622	0.011	0.618	0.010	0.741	0.004
spa	XLM-RoBERTa (base)	0.632	0.012	0.617	0.011	0.730	0.009
glg	BERTimbau (base)	0.606	0.013	0.613	0.013	0.744	0.006
glg	IXAes	0.611	0.012	0.609	0.010	0.735	0.006
spa	Bertinho	0.599	0.010	0.605	0.011	0.738	0.006
spa	BERTimbau (base)	0.604	0.012	0.595	0.011	0.725	0.006
eng	BERT multilingual (base)	0.578	0.013	0.587	0.011	0.727	0.005
glg	BERT multilingual (base)	0.570	0.013	0.579	0.010	0.721	0.005
spa	BERT multilingual (base)	0.568	0.009	0.569	0.009	0.711	0.006
eng	IXAes	0.563	0.013	0.556	0.011	0.698	0.005
por	BERT (large)	0.548	0.112	0.554	0.136	0.731	0.038
eng	BERTimbau (base)	0.524	0.012	0.539	0.013	0.702	0.010
eng	Bertinho	0.526	0.007	0.537	0.010	0.698	0.006
eng	BERTimbau (large)	0.491	0.128	0.490	0.157	0.695	0.034

Table B.8: Complete version: ReLi-SA results.

B.1.6 | PorSimplesSent

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	mDeBERTa v3 (base)	0.952	0.003	0.953	0.003	0.957	0.003
por	BERT multilingual (base)	0.932	0.005	0.933	0.005	0.938	0.004
por	XLM-RoBERTa (base)	0.928	0.006	0.929	0.006	0.934	0.005
por	BERTimbau (base)	0.918	0.013	0.920	0.013	0.926	0.012
por	BERTimbau (large)	0.918	0.012	0.919	0.012	0.926	0.011
por	BERT (large)	0.906	0.005	0.907	0.006	0.915	0.005
por	Bertinho	0.898	0.003	0.900	0.003	0.908	0.003
por	IXAes	0.899	0.008	0.899	0.008	0.908	0.007
glg	mDeBERTa v3 (base)	0.875	0.005	0.878	0.005	0.887	0.004
glg	BERT multilingual (base)	0.869	0.003	0.873	0.003	0.880	0.003
spa	mDeBERTa v3 (base)	0.863	0.009	0.869	0.010	0.876	0.008
glg	XLM-RoBERTa (base)	0.860	0.004	0.864	0.004	0.873	0.003
spa	BERT multilingual (base)	0.856	0.006	0.862	0.006	0.869	0.005
eng	mDeBERTa v3 (base)	0.853	0.002	0.859	0.002	0.864	0.002
glg	BERTimbau (large)	0.852	0.005	0.855	0.006	0.866	0.005
glg	BERTimbau (base)	0.849	0.006	0.852	0.006	0.862	0.005
spa	XLM-RoBERTa (base)	0.846	0.004	0.852	0.005	0.861	0.004
spa	BERTimbau (base)	0.844	0.005	0.849	0.005	0.858	0.005
spa	BERT (large)	0.840	0.004	0.845	0.004	0.854	0.004
eng	BERT multilingual (base)	0.838	0.004	0.844	0.004	0.851	0.004
spa	Bertinho	0.833	0.004	0.839	0.005	0.849	0.004
glg	IXAes	0.834	0.007	0.837	0.007	0.851	0.006
eng	XLM-RoBERTa (base)	0.830	0.008	0.835	0.008	0.844	0.007
eng	BERTimbau (base)	0.829	0.007	0.833	0.007	0.843	0.006
glg	Bertinho	0.825	0.008	0.828	0.009	0.842	0.007
spa	IXAes	0.822	0.011	0.825	0.011	0.838	0.010
eng	IXAes	0.816	0.007	0.820	0.007	0.833	0.006
eng	Bertinho	0.810	0.005	0.814	0.004	0.827	0.004
spa	BERTimbau (large)	0.791	0.161	0.781	0.208	0.811	0.146
eng	BERT (large)	0.686	0.244	0.647	0.316	0.714	0.220
eng	BERTimbau (large)	0.682	0.241	0.642	0.313	0.711	0.218
glg	BERT (large)	0.434	0.213	0.319	0.275	0.487	0.193

Table B.9: Complete version: PorSimplesSent results.

B.1.7 | FaQuaD-NLI

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.907	0.003	0.900	0.005	0.931	0.004
por	BERTimbau (base)	0.898	0.011	0.897	0.011	0.930	0.007
spa	mDeBERTa v3 (base)	0.898	0.010	0.889	0.008	0.923	0.006
por	mDeBERTa v3 (base)	0.888	0.011	0.887	0.013	0.923	0.009
eng	mDeBERTa v3 (base)	0.891	0.008	0.883	0.011	0.919	0.009
glg	mDeBERTa v3 (base)	0.877	0.013	0.869	0.013	0.910	0.009
glg	BERTimbau (large)	0.874	0.009	0.867	0.009	0.908	0.007
por	Bertinho	0.878	0.005	0.866	0.004	0.907	0.003
por	BERT multilingual (base)	0.863	0.009	0.865	0.012	0.909	0.009
spa	BERT multilingual (base)	0.863	0.009	0.865	0.006	0.909	0.004
por	IXAes	0.870	0.009	0.860	0.010	0.903	0.007
glg	Bertinho	0.868	0.011	0.858	0.009	0.901	0.006
glg	BERTimbau (base)	0.862	0.006	0.857	0.006	0.902	0.005
eng	BERT multilingual (base)	0.853	0.011	0.855	0.009	0.902	0.007
spa	BERTimbau (large)	0.855	0.017	0.854	0.013	0.901	0.009
spa	IXAes	0.864	0.011	0.854	0.010	0.899	0.007
glg	BERT multilingual (base)	0.857	0.014	0.853	0.015	0.899	0.010
spa	BERTimbau (base)	0.853	0.008	0.850	0.007	0.898	0.005
spa	Bertinho	0.862	0.010	0.849	0.008	0.894	0.006
spa	XLM-RoBERTa (base)	0.853	0.009	0.843	0.011	0.891	0.009
eng	XLM-RoBERTa (base)	0.852	0.010	0.841	0.015	0.889	0.013
spa	BERT (large)	0.840	0.006	0.838	0.006	0.890	0.006
eng	BERTimbau (large)	0.848	0.009	0.838	0.008	0.888	0.006
glg	BERT (large)	0.840	0.008	0.834	0.010	0.886	0.010
glg	IXAes	0.851	0.006	0.831	0.012	0.879	0.011
eng	BERTimbau (base)	0.837	0.010	0.829	0.010	0.882	0.008
eng	Bertinho	0.835	0.007	0.827	0.008	0.880	0.007
eng	IXAes	0.842	0.012	0.825	0.012	0.876	0.009
por	XLM-RoBERTa (base)	0.794	0.156	0.776	0.178	0.879	0.052
glg	XLM-RoBERTa (base)	0.783	0.126	0.766	0.133	0.860	0.036
eng	BERT (large)	0.679	0.189	0.644	0.216	0.840	0.059
por	BERT (large)	0.602	0.164	0.559	0.192	0.816	0.050

Table B.10: Complete version: FaQuaD-NLI results.

B.1.8 | HateBR

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.919	0.005	0.919	0.005	0.919	0.005
por	BERTimbau (base)	0.914	0.003	0.913	0.003	0.914	0.003
por	mDeBERTa v3 (base)	0.911	0.004	0.911	0.004	0.911	0.004
por	XLM-RoBERTa (base)	0.902	0.004	0.902	0.004	0.902	0.004
por	Bertinho	0.879	0.005	0.879	0.005	0.879	0.005
glg	mDeBERTa v3 (base)	0.873	0.004	0.873	0.004	0.873	0.004
por	IXAes	0.872	0.005	0.872	0.005	0.872	0.005
por	BERT multilingual (base)	0.871	0.007	0.871	0.007	0.871	0.007
glg	Bertinho	0.870	0.007	0.870	0.007	0.870	0.007
spa	mDeBERTa v3 (base)	0.870	0.008	0.870	0.008	0.870	0.008
spa	IXAes	0.868	0.005	0.868	0.005	0.868	0.005
glg	BERTimbau (base)	0.867	0.005	0.867	0.005	0.867	0.005
spa	XLM-RoBERTa (base)	0.865	0.006	0.865	0.006	0.865	0.006
spa	BERTimbau (base)	0.858	0.004	0.858	0.004	0.858	0.004
eng	mDeBERTa v3 (base)	0.857	0.004	0.857	0.004	0.857	0.004
spa	Bertinho	0.854	0.006	0.854	0.006	0.854	0.006
eng	XLM-RoBERTa (base)	0.852	0.009	0.852	0.009	0.852	0.009
eng	BERT multilingual (base)	0.847	0.007	0.847	0.007	0.847	0.007
spa	BERT multilingual (base)	0.846	0.005	0.846	0.005	0.846	0.005
glg	BERT multilingual (base)	0.838	0.003	0.838	0.003	0.838	0.003
por	BERT (large)	0.838	0.054	0.838	0.055	0.838	0.054
glg	IXAes	0.833	0.084	0.832	0.087	0.833	0.084
eng	IXAes	0.832	0.005	0.832	0.005	0.832	0.005
eng	Bertinho	0.819	0.009	0.819	0.009	0.819	0.009
eng	BERTimbau (large)	0.818	0.005	0.818	0.005	0.818	0.005
glg	BERTimbau (large)	0.831	0.116	0.814	0.169	0.831	0.116
glg	BERT (large)	0.811	0.042	0.810	0.044	0.811	0.042
eng	BERT (large)	0.825	0.114	0.809	0.167	0.826	0.114
eng	BERTimbau (base)	0.808	0.009	0.808	0.009	0.808	0.009
glg	XLM-RoBERTa (base)	0.812	0.113	0.796	0.165	0.812	0.113
spa	BERTimbau (large)	0.642	0.183	0.542	0.269	0.642	0.183
spa	BERT (large)	0.613	0.145	0.520	0.224	0.613	0.145

Table B.11: Complete version: HateBR results.

B.1.9 | ReRelEM

language	model	Recall		F1 Score		Accuracy	
		mean	std	mean	std	mean	std
por	BERTimbau (large)	0.316	0.029	0.300	0.032	0.648	0.013
por	BERTimbau (base)	0.260	0.034	0.247	0.033	0.607	0.012
por	BERT multilingual (base)	0.193	0.025	0.183	0.022	0.550	0.026
por	Bertinho	0.189	0.025	0.175	0.024	0.580	0.012
por	mDeBERTa v3 (base)	0.153	0.021	0.150	0.020	0.582	0.014
por	IXAes	0.125	0.009	0.117	0.011	0.539	0.013
por	BERT (large)	0.116	0.007	0.113	0.007	0.535	0.016
por	XLM-RoBERTa (base)	0.112	0.011	0.098	0.012	0.530	0.023

Table B.12: Complete version: ReRelEM results.

B.2 | Fine-Grained Evaluation

B.2.1 | Evaluation Statistics

To facilitate the reproducibility of our fine-grained evaluation, we provide statistics for the buckets defined for each attribute and dataset. Here is the precise meaning of each column in the tables below:

- **attribute**: the attribute name.
- **bucket**: the bucket number, from 0 to 3.
- **start**: the start of the range, inclusive, for the bucket.
- **end**: the end of the range, exclusive, for the bucket.
- **samples**: the number of samples in the bucket.

For each attribute, there are four buckets, numbered from 0 to 3. Each bucket is characterized by a range of values, with all samples within that range assigned to the respective bucket. The ranges for most attributes are given in percentages. However, for the len attribute, they are indicated by character length. A few examples will clarify:

- In Table B.13, the third bucket under the len attribute is defined as the range from 249.25 to 313 characters. Hence, only samples with length in the interval $[250, 313[$ are included.

- In Table B.14, the second bucket under the R_{OOV} attribute is defined as the range from 0.167 to 0.25. Thus, all samples with a R_{OOV} value in the interval $[16.7\%, 25\%[$ are assigned to this bucket.

In both Table B.18 and Table B.15, the R_{WO} attribute is notably absent. This is because this attribute is applicable only to datasets based on sentence pairs, not to those with single sentences.

Our goal was to have equal interval ranges for each bucket. Therefore, while the intervals remain consistent, the number of samples in each bucket might vary. The range of the last bucket always includes any samples that did not fit into the earlier buckets. This ensures that every sample is assigned to a bucket, with none being left out.

B.2.1.1 | ASSIN

attribute	bucket	start	end	samples
R_{WO}	0	0.03	0.138	764
	1	0.138	0.246	2321
	2	0.246	0.354	841
	3	0.354	0.462	73
R_{OOV}	0	0.0	0.098	2632
	1	0.098	0.197	1061
	2	0.197	0.295	278
	3	0.295	0.394	28
F_{train}	0	0.071	0.125	256
	1	0.125	0.178	2457
	2	0.178	0.232	1234
	3	0.232	0.286	52
len	0	58.0	121.75	1310
	1	121.75	185.5	2544
	2	185.5	249.25	145
	3	249.25	313.0	0
LC - Entailment	0	0.027	0.227	845
	1	0.227	0.427	217
	2	0.427	0.627	724
	3	0.627	0.827	2213
LC - Similarity	0	0.014	0.152	705
	1	0.152	0.289	1300
	2	0.289	0.426	1901
	3	0.426	0.564	93

Table B.13: Fine-grained evaluation details for ASSIN.

B.2.1.2 | ASSIN 2

attribute	bucket	start	end	samples
R_{WO}	0	0.0	0.125	297
	1	0.125	0.25	681
	2	0.25	0.375	946
	3	0.375	0.5	521
R_{OOV}	0	0.0	0.083	2351
	1	0.083	0.167	73
	2	0.167	0.25	21
	3	0.25	0.333	2
F_{train}	0	0.031	0.137	276
	1	0.137	0.243	1271
	2	0.243	0.348	866
	3	0.348	0.454	34
len	0	35.0	91.25	1440
	1	91.25	147.5	891
	2	147.5	203.75	105
	3	203.75	260.0	11
LC - Entailment	0	0.262	0.389	267
	1	0.389	0.517	1052
	2	0.517	0.644	1033
	3	0.644	0.771	95
LC - Similarity	0	0.018	0.21	825
	1	0.21	0.402	716
	2	0.402	0.593	675
	3	0.593	0.785	231

Table B.14: Fine-grained evaluation details for ASSIN 2.

B.2.1.3 | ReLi-SA

attribute	bucket	start	end	samples
R_{OOV}	0	0.0	0.25	2865
	1	0.25	0.5	228
	2	0.5	0.75	90
	3	0.75	1.0	9
F_{train}	0	0.0	0.127	518
	1	0.127	0.254	1579
	2	0.254	0.382	1121
	3	0.382	0.509	69
len	0	1.0	327.75	3267
	1	327.75	654.5	17
	2	654.5	981.25	3
	3	981.25	1308.0	0
LC - Task Label	0	0.0	0.25	1076
	1	0.25	0.5	463
	2	0.5	0.75	1708
	3	0.75	1.0	31

Table B.15: Fine-grained evaluation details for ReLi-SA.

B.2.1.4 | PorSimplesSent

attribute	bucket	start	end	samples
R_{WO}	0	0.0	0.125	6
	1	0.125	0.25	110
	2	0.25	0.375	505
	3	0.375	0.5	616
R_{OOV}	0	0.0	0.167	947
	1	0.167	0.333	638
	2	0.333	0.5	99
	3	0.5	0.667	11
F_{train}	0	0.012	0.107	27
	1	0.107	0.201	905
	2	0.201	0.295	723
	3	0.295	0.389	40
len	0	18.0	145.0	843
	1	145.0	272.0	649
	2	272.0	399.0	186
	3	399.0	526.0	18
LC - Task Label	0	0.096	0.211	170
	1	0.211	0.327	1160
	2	0.327	0.443	357
	3	0.443	0.558	9

Table B.16: Fine-grained evaluation details for PorSimplesSent.

B.2.1.5 | FaQuaD-NLI

attribute	bucket	start	end	samples
R_{WO}	0	0.0	0.075	376
	1	0.075	0.15	204
	2	0.15	0.225	49
	3	0.225	0.3	20
R_{OOV}	0	0.0	0.089	221
	1	0.089	0.179	329
	2	0.179	0.268	86
	3	0.268	0.357	13
F_{train}	0	0.064	0.135	89
	1	0.135	0.206	359
	2	0.206	0.276	187
	3	0.276	0.347	14
len	0	55.0	180.0	240
	1	180.0	305.0	339
	2	305.0	430.0	54
	3	430.0	555.0	16
LC - Task Label	0	0.149	0.312	139
	1	0.312	0.475	1
	2	0.475	0.638	161
	3	0.638	0.801	348

Table B.17: Fine-grained evaluation details for FaQuaD-NLI.

B.2.1.6 | HateBR

attribute	bucket	start	end	samples
R_{OOV}	0	0.0	0.25	1198
	1	0.25	0.5	136
	2	0.5	0.75	42
	3	0.75	1.0	3
F_{train}	0	0.0	0.214	1064
	1	0.214	0.429	278
	2	0.429	0.643	46
	3	0.643	0.857	11
len	0	3.0	164.5	1335
	1	164.5	326.0	48
	2	326.0	487.5	12
	3	487.5	649.0	4
LC - Offensive Language	0	0.0	0.25	47
	1	0.25	0.5	521
	2	0.5	0.75	627
	3	0.75	1.0	162

Table B.18: Fine-grained evaluation details for HateBR.

B.2.1.7 | ReRelEM

attribute	bucket	start	end	samples
R_{WO}	0	0.0	0.057	455
	1	0.057	0.114	240
	2	0.114	0.171	80
	3	0.171	0.228	27
R_{OOV}	0	0.0	0.156	115
	1	0.156	0.312	628
	2	0.312	0.469	55
	3	0.469	0.625	6
F_{train}	0	0.053	0.106	64
	1	0.106	0.16	402
	2	0.16	0.213	329
	3	0.213	0.266	9
len	0	42.0	742.25	558
	1	742.25	1442.5	153
	2	1442.5	2142.75	62
	3	2142.75	2843.0	24
LC - Task Label	0	0.0	0.132	387
	1	0.132	0.263	4
	2	0.263	0.394	160
	3	0.394	0.526	253

Table B.19: Fine-grained evaluation details for ReRelEM.

B.2.2 | Full Results

This section presents the full results of our fine-grained evaluation. Each table in this section corresponds to a different attribute:

- len (Table B.20)
- *LC* (Table B.21)
- R_{OOV} (Table B.22)
- R_{WO} (Table B.23)
- F_{train} (Table B.24)

model	dataset bucket	assin-rte	assin-sts	assin2-rte	assin2-sts	faquad	hatebr	pss	reli-sa	rerelem
BERT multilingual (base)	0	0.82	0.807	0.885	0.845	0.859	0.873	0.906	0.643	0.205
	1	0.804	0.819	0.844	0.770	0.879	0.793	0.941	0.572	0.138
	2	0.844	0.851	0.836	0.728	0.770	0.816	0.907	0.244	0.112
	3	-	-	0.901	0.874	0.982	0.960	0.987	0.100	0.095
BERTimbau (base)	0	0.846	0.836	0.900	0.870	0.889	0.916	0.881	0.714	0.275
	1	0.818	0.84	0.858	0.788	0.912	0.832	0.928	0.645	0.183
	2	0.833	0.879	0.860	0.774	0.816	0.891	0.917	0.200	0.113
	3	-	-	0.901	0.833	1.000	0.980	0.949	0.400	0.133
BERTimbau (large)	0	0.855	0.756	0.912	0.882	0.893	0.920	0.879	0.746	0.325
	1	0.829	0.757	0.876	0.811	0.913	0.867	0.930	0.736	0.229
	2	0.839	0.796	0.864	0.782	0.826	0.908	0.913	0.311	0.130
	3	-	-	0.860	0.878	0.988	0.929	0.973	0.300	0.142
Bertinho	0	0.775	0.754	0.870	0.822	0.853	0.883	0.869	0.680	0.203
	1	0.758	0.779	0.813	0.738	0.880	0.754	0.903	0.692	0.146
	2	0.764	0.841	0.804	0.729	0.806	0.807	0.877	0.222	0.141
	3	-	-	0.891	0.869	0.872	0.823	0.913	0.700	0.119
IXAes	0	0.761	0.74	0.864	0.795	0.841	0.874	0.864	0.637	0.130
	1	0.748	0.765	0.802	0.698	0.879	0.803	0.912	0.636	0.126
	2	0.825	0.827	0.761	0.665	0.797	0.790	0.900	0.222	0.150
	3	-	-	0.832	0.830	0.817	0.810	0.805	0.000	0.127
XLM-RoBERTa (base)	0	0.831	0.801	0.756	0.852	0.766	0.905	0.895	0.703	0.115
	1	0.816	0.808	0.742	0.756	0.791	0.806	0.941	0.592	0.097
	2	0.836	0.864	0.726	0.750	0.689	0.864	0.898	0.233	0.084
	3	-	-	0.737	0.852	0.870	0.833	0.971	0.900	0.141
mDeBERTa v3 (base)	0	0.876	0.849	0.920	0.876	0.875	0.915	0.933	0.719	0.172
	1	0.856	0.851	0.880	0.798	0.907	0.830	0.962	0.638	0.121
	2	0.867	0.89	0.860	0.764	0.803	0.825	0.936	0.244	0.119
	3	-	-	0.902	0.862	0.797	0.744	0.794	0.900	0.102

Table B.20: Full fine-grained evaluation results: len results for each bucket and task.

model	dataset bucket	assin-rte	assin-sts	assin2-rte	assin2-sts	faquad	hatebr	pss	reli-sa	rerelem
BERT multilingual (base)	0	0.522	0.903	0.428	0.850	0.438	0.798	0.855	0.527	0.176
	1	0.318	0.749	0.808	0.322	1.000	0.771	0.929	0.620	0.707
	2	0.324	0.618	0.921	0.398	0.487	0.920	0.949	0.339	0.061
	3	0.326	0.548	0.895	0.419	0.485	0.968	0.852	0.930	0.071
BERTimbau (base)	0	0.529	0.915	0.445	0.864	0.457	0.928	0.826	0.631	0.241
	1	0.341	0.781	0.823	0.345	1.000	0.852	0.917	0.646	0.600
	2	0.325	0.660	0.945	0.438	0.495	0.939	0.932	0.346	0.068
	3	0.328	0.507	0.968	0.478	0.485	0.974	0.731	0.907	0.077
BERTimbau (large)	0	0.539	0.823	0.468	0.873	0.464	0.936	0.819	0.684	0.293
	1	0.346	0.705	0.839	0.368	1.000	0.855	0.915	0.639	0.607
	2	0.325	0.608	0.957	0.457	0.491	0.948	0.939	0.330	0.093
	3	0.328	0.474	0.965	0.487	0.485	0.971	0.727	0.889	0.079
Bertinho	0	0.484	0.863	0.400	0.813	0.452	0.853	0.814	0.590	0.166
	1	0.329	0.700	0.774	0.306	1.000	0.776	0.899	0.626	0.600
	2	0.319	0.582	0.920	0.359	0.485	0.931	0.906	0.341	0.094
	3	0.324	0.407	0.979	0.417	0.480	0.966	0.720	0.934	0.100
IXAes	0	0.479	0.864	0.391	0.780	0.447	0.806	0.815	0.525	0.104
	1	0.306	0.673	0.759	0.271	1.000	0.767	0.895	0.653	0.693
	2	0.321	0.557	0.911	0.338	0.487	0.916	0.920	0.345	0.118
	3	0.325	0.445	0.908	0.323	0.479	0.992	0.751	0.915	0.097
XLM-RoBERTa (base)	0	0.534	0.895	0.516	0.841	0.357	0.790	0.820	0.617	0.081
	1	0.333	0.748	0.730	0.339	0.800	0.811	0.929	0.651	0.740
	2	0.323	0.604	0.718	0.402	0.587	0.947	0.943	0.331	0.102
	3	0.326	0.538	0.674	0.440	0.585	0.993	0.793	0.958	0.130
mDeBERTa v3 (base)	0	0.566	0.928	0.492	0.863	0.452	0.861	0.921	0.649	0.135
	1	0.343	0.801	0.846	0.376	1.000	0.828	0.951	0.641	0.600
	2	0.325	0.672	0.951	0.445	0.489	0.951	0.960	0.332	0.083
	3	0.328	0.559	0.976	0.440	0.486	0.994	0.945	0.823	0.108

Table B.21: Full fine-grained evaluation results: *LC* results for each bucket and task.

model	dataset bucket	assin-rte	assin-sts	assin2-rte	assin2-sts	faquad	hatebr	pss	reli-sa	rerelem
BERT multilingual (base)	0	0.807	0.823	0.870	0.818	0.836	0.880	0.946	0.649	0.312
	1	0.825	0.817	0.857	0.801	0.868	0.802	0.920	0.489	0.185
	2	0.810	0.802	0.711	0.692	0.902	0.834	0.804	0.482	0.224
	3	0.899	0.877	0.690	-0.606	0.976	0.755	0.587	0.550	0.500
BERTimbau (base)	0	0.826	0.847	0.885	0.841	0.852	0.916	0.930	0.719	0.420
	1	0.836	0.843	0.866	0.818	0.910	0.866	0.907	0.540	0.230
	2	0.807	0.824	0.796	0.727	0.951	0.949	0.797	0.443	0.259
	3	0.964	0.860	0.732	-0.958	0.963	0.914	0.594	0.636	0.446
BERTimbau (large)	0	0.838	0.764	0.898	0.856	0.868	0.920	0.930	0.752	0.437
	1	0.841	0.758	0.875	0.860	0.907	0.885	0.904	0.574	0.295
	2	0.815	0.742	0.831	0.846	0.943	0.934	0.820	0.568	0.278
	3	0.882	0.753	0.920	-0.728	0.988	0.919	0.676	0.636	0.415
Bertinho	0	0.766	0.782	0.849	0.793	0.829	0.886	0.910	0.687	0.367
	1	0.766	0.777	0.825	0.777	0.876	0.827	0.892	0.543	0.183
	2	0.747	0.756	0.781	0.618	0.885	0.851	0.747	0.441	0.213
	3	0.827	0.793	0.860	-0.365	1.000	0.819	0.701	0.621	0.470
IXAes	0	0.755	0.764	0.840	0.760	0.795	0.878	0.907	0.642	0.246
	1	0.772	0.769	0.811	0.769	0.885	0.821	0.893	0.520	0.128
	2	0.741	0.748	0.669	0.823	0.895	0.875	0.794	0.600	0.177
	3	0.622	0.783	0.388	-0.868	1.000	0.781	0.533	0.603	0.319
XLM-RoBERTa (base)	0	0.815	0.815	0.753	0.819	0.738	0.909	0.942	0.708	0.231
	1	0.844	0.807	0.708	0.796	0.785	0.851	0.917	0.540	0.102
	2	0.804	0.804	0.666	0.750	0.815	0.889	0.817	0.580	0.161
	3	0.952	0.884	0.675	-0.299	0.859	0.796	0.612	0.618	0.559
mDeBERTa v3 (base)	0	0.860	0.860	0.904	0.847	0.850	0.918	0.957	0.725	0.322
	1	0.876	0.849	0.901	0.865	0.893	0.858	0.948	0.625	0.153
	2	0.838	0.839	0.887	0.780	0.930	0.848	0.878	0.601	0.191
	3	0.859	0.842	0.860	-0.050	0.988	0.922	0.771	0.596	0.315

Table B.22: Full fine-grained evaluation results: R_{OOV} results for each bucket and task.

model	dataset bucket	assin-rte	assin-sts	assin2-rte	assin2-sts	faquad	hatebr	pss	reli-sa	rerelem
BERT multilingual (base)	0	0.568	0.604	0.848	0.762	0.772	0.871	0.674	0.642	0.202
	1	0.747	0.740	0.860	0.798	0.841	0.871	0.977	0.642	0.171
	2	0.802	0.763	0.853	0.672	0.786	0.871	0.974	0.642	0.288
	3	0.656	0.692	0.848	0.660	0.754	0.871	0.870	0.642	0.358
BERTimbau (base)	0	0.657	0.656	0.885	0.826	0.815	0.913	0.690	0.713	0.274
	1	0.778	0.774	0.882	0.819	0.886	0.913	0.979	0.713	0.219
	2	0.814	0.794	0.861	0.704	0.835	0.913	0.869	0.713	0.260
	3	0.739	0.759	0.865	0.691	0.851	0.913	0.852	0.713	0.304
BERTimbau (large)	0	0.668	0.615	0.897	0.839	0.817	0.919	0.567	0.745	0.312
	1	0.788	0.705	0.895	0.839	0.897	0.919	0.976	0.745	0.301
	2	0.824	0.720	0.880	0.738	0.795	0.919	0.850	0.745	0.311
	3	0.730	0.744	0.883	0.715	0.882	0.919	0.854	0.745	0.459
Bertinho	0	0.500	0.512	0.820	0.719	0.793	0.879	0.430	0.681	0.188
	1	0.654	0.669	0.840	0.766	0.852	0.879	0.973	0.681	0.233
	2	0.762	0.706	0.827	0.659	0.686	0.879	0.846	0.681	0.278
	3	0.638	0.630	0.828	0.631	0.624	0.879	0.819	0.681	0.443
IXAes	0	0.625	0.500	0.826	0.654	0.757	0.872	0.635	0.637	0.132
	1	0.664	0.652	0.822	0.727	0.849	0.872	0.953	0.637	0.129
	2	0.751	0.668	0.818	0.617	0.765	0.872	0.885	0.637	0.214
	3	0.643	0.658	0.823	0.632	0.621	0.872	0.821	0.637	0.170
XLM-RoBERTa (base)	0	0.619	0.562	0.759	0.793	0.711	0.902	0.675	0.702	0.108
	1	0.758	0.728	0.750	0.800	0.752	0.902	0.983	0.702	0.131
	2	0.815	0.764	0.730	0.676	0.668	0.902	0.844	0.702	0.184
	3	0.657	0.765	0.719	0.669	0.572	0.902	0.868	0.702	0.223
mDeBERTa v3 (base)	0	0.780	0.672	0.909	0.834	0.828	0.911	0.657	0.719	0.160
	1	0.836	0.791	0.892	0.824	0.854	0.911	0.987	0.719	0.187
	2	0.841	0.821	0.886	0.730	0.848	0.911	0.981	0.719	0.207
	3	0.772	0.824	0.908	0.750	0.873	0.911	0.905	0.719	0.397

Table B.23: Full fine-grained evaluation results: R_{WO} results for each bucket and task.

model	dataset bucket	assin-rte	assin-sts	assin2-rte	assin2-sts	faquad	hatebr	pss	reli-sa	rerelem
BERT multilingual (base)	0	0.849	0.861	0.847	0.706	0.836	0.864	0.729	0.655	0.290
	1	0.816	0.818	0.873	0.797	0.853	0.872	0.932	0.620	0.172
	2	0.784	0.814	0.866	0.859	0.897	0.982	0.936	0.664	0.177
	3	0.566	0.791	0.894	0.813	1.000	0.991	0.952	0.532	0.634
BERTimbau (base)	0	0.863	0.878	0.879	0.737	0.876	0.909	0.726	0.732	0.313
	1	0.829	0.845	0.886	0.817	0.896	0.921	0.913	0.693	0.182
	2	0.807	0.831	0.879	0.881	0.903	0.939	0.930	0.733	0.263
	3	0.590	0.823	0.904	0.888	1.000	1.000	0.953	0.568	0.570
BERTimbau (large)	0	0.872	0.790	0.898	0.771	0.882	0.913	0.778	0.770	0.380
	1	0.832	0.762	0.900	0.832	0.900	0.928	0.914	0.717	0.249
	2	0.833	0.747	0.890	0.893	0.898	0.961	0.927	0.778	0.305
	3	0.647	0.746	0.929	0.926	0.991	1.000	0.784	0.594	0.760
Bertinho	0	0.781	0.832	0.822	0.687	0.869	0.872	0.722	0.697	0.257
	1	0.771	0.778	0.852	0.768	0.856	0.888	0.893	0.660	0.195
	2	0.739	0.762	0.846	0.837	0.874	0.972	0.910	0.701	0.189
	3	0.507	0.770	0.863	0.787	1.000	0.920	0.961	0.599	0.669
IXAes	0	0.782	0.828	0.814	0.638	0.839	0.870	0.682	0.639	0.217
	1	0.761	0.764	0.841	0.747	0.860	0.863	0.895	0.616	0.126
	2	0.736	0.746	0.836	0.801	0.857	0.965	0.906	0.655	0.128
	3	0.573	0.745	0.869	0.688	1.000	0.837	0.909	0.558	0.689
XLM-RoBERTa (base)	0	0.864	0.856	0.743	0.708	0.776	0.895	0.715	0.724	0.253
	1	0.823	0.812	0.759	0.795	0.771	0.915	0.931	0.675	0.111
	2	0.803	0.802	0.738	0.861	0.778	0.951	0.931	0.735	0.101
	3	0.579	0.774	0.806	0.818	0.872	0.991	0.934	0.578	0.418
mDeBERTa v3 (base)	0	0.877	0.889	0.902	0.750	0.909	0.910	0.853	0.716	0.243
	1	0.868	0.853	0.904	0.822	0.878	0.907	0.950	0.705	0.159
	2	0.848	0.851	0.902	0.889	0.881	0.949	0.958	0.739	0.157
	3	0.606	0.829	0.914	0.878	0.981	0.947	0.938	0.558	0.481

Table B.24: Full fine-grained evaluation results: F_{train} results for each bucket and task.

References

Rodrigo Agerri and Eneko Agirre. Lessons learned from the evaluation of spanish language models. *arXiv preprint arXiv:2212.08390*, 2022.

Sandra Aluísio, Jorge Pelizzoni, Ana Raquel Marchi, Lucélia de Oliveira, Regiana Manenti, and Vanessa Marquiafável. An account of the challenge of tagging a reference corpus for brazilian portuguese. In *International workshop on computational processing of the portuguese language*, pages 110–117. Springer, 2003.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.618. URL `https://doi.org/10.18653%2Fv1%2F2020.emnlp-main.618`.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. Does corpus quality really matter for low-resource languages? *arXiv preprint arXiv:2203.08111*, 2022.

Jack Bandy and Nicholas Vincent. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Enrique Bigne, Carla Ruiz, Carmen Perez-Cabañero, and Antonio Cuenca. Are customer star ratings and sentiments aligned? a deep learning study of the customer service experience in tourism destinations. *Service Business*, 17(1):281–314, 2023.

Henrico Bertini Brum and Maria das Graças Volpe Nunes. Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*, 2017.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

Isabel Carvalho, Hugo Gonçalo Oliveira, and Catarina Silva. Sentiment Analysis in Portuguese Dialogues . In *Proc. IberSPEECH 2022*, pages 176–180, 2022. doi: 10.21437/IberSPEECH.2022-36.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef van Genabith. Towards debiasing translation artifacts. *arXiv preprint arXiv:2205.08001*, 2022.

Kenneth Ward Church and Valia Kordoni. Emerging trends: Sota-chasing. *Natural Language Engineering*, 28(2):249–269, 2022.

Sandra Collovini, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. In *IberLEF@ SEPLN*, pages 390–410, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

Ulisses B Corrêa, Leonardo Coelho, Leonardo Santos, and Larissa A de Freitas. Overview of the idpt task on irony detection in portuguese at iberlef 2021. *Procesamiento del Lenguaje Natural*, 67: 269–276, 2021.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Lucas Cabral Carneiro da Cunha. Fakewhatsapp. br: detecção de desinformação e desinformadores em grupos públicos do whatsapp em pt-br. 2021.

Felix LV da Silva, Guilherme da S Xavier, Heliks M Mensenburg, Rodrigo F Rodrigues, Leonardo P dos Santos, Ricardo M Araújo, Ulisses B Corrêa, and Larissa A de Freitas. Absapt 2022 at iberlef: Overview of the task on aspect-based sentiment analysis in portuguese. *Procesamiento del Lenguaje Natural*, 69:199–205, 2022.

Pedro Vitor Quinta de Castro, Nádia Félix Felipe da Silva, and Anderson da Silva Soares. Contextual representations and semi-supervised named entity recognition for portuguese language. In *IberLEF SEPLN*, pages 411–420, 2019.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer, 2018.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

Jacob Devlin. Bert. `https://github.com/google-research/bert`, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392, 2018.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance-how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, 2019.

Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. Do we name the languages we study? the# benderrule in lrec and acl articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 564–573, 2022.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*, 2019.

Diogo Fernandes, Adalberto Junior, Gabriel Marques, Anderson Soares, and Arlindo Galvao Filho. Ceia-nlp at case 2022 task 1: Protest news detection for portuguese. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 184–188, 2022.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15, 2016.

Erick Rocha Fonseca and Sandra Maria Aluísio. Improving pos tagging across portuguese variants with word embeddings. In *International Conference on Computational Processing of the Portuguese Language*, pages 227–232. Springer, 2016.

Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Goncalo Oliveira, and Paula Carvalho. Relation detection between named entities: report of a shared task. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 129–137, 2009.

Cláudia Freitas, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota, and Diana Santos. Second harem: advancing the state of the art of named entity recognition in portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association*. European Language Resources Association, 2010.

Cláudia Freitas, Eduardo Motta, Ruy Luiz Milidiú, and Juliana César. Sparkling vampire... lol! annotating opinions in a book review corpus. In Sandra Aluísio and Stella E. O. Tagnin, editors, *New Language Technologies and Linguistic Research: A Two-Way Road*, pages 128–146. Cambridge Scholars Publishing, 2014.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. Rethinkcws: Is chinese word segmentation a solved task? *arXiv preprint arXiv:2011.06858*, 2020.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.478`.

Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. Deep learning tuning playbook, 2023. URL `http://github.com/google-research/tuning_playbook`. Version 1.0.

JRS Gomes. Plue: Portuguese language understanding evaluation. `https://github.com/ju-resplande/PLUE`, 2020.

JRS Gomes, RC Rodrigues, EAS Garcia, AFB Junior, Diogo Fernandes Costa Silva, and Dyonnatan Ferreira Maia. 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*, 2022.

Hugo Gonçalo Oliveira, João Ferreira, José Santos, Pedro Fialho, Ricardo Rodrigues, Luísa Coheur, and Ana Alves. AIA-BDE: A corpus of faqs in portuguese and their variations. In *Proceedings of 12th International Conference on Language Resources and Evaluation*, LREC 2020, pages 5442—5449, Marseille, France, 2020. ELRA.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of bert. *arXiv preprint arXiv:1908.05620*, 2019.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.

Nathan Siegle Hartmann and Sandra Maria Aluísio. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. *Linguamática*, 12(2):3–27, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Selçuk Gürel, Benjamin Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, et al. Extended multilingual protest news detection–shared task 1, case 2021 and 2022. *arXiv preprint arXiv:2211.11360*, 2022.

LR Iman and JM Davenport. Approximations of the critical region of the friedman statistic. *Communications in Statics*, pages 571–595, 1980.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. Cabrita: closing the gap for foreign languages, 2023.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413, Santa Fe, New Mexico, USA, August 2018a. Association for Computational Linguistics. URL `https://aclanthology.org/C18-1034`.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluíso. A nontrivial sentence corpus for the task of sentence readability assessment in portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 401–413, Santa Fe, New Mexico, USA, August 2018b.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. Lener-br: a dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 313–323. Springer, 2018.

Marco Marelli, Stefano Menini, Marco Baroni, and Luisa Bentivogli. Raffaella bernardi, and roberto zamparelli. 2014. a sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

João Moreno and Graça Bressan. Factck. br: a new dataset to study fake news. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 525–527, 2019.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.

Cristina Mota and Diana Santos. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem, 2008.

Susan M Mudambi, David Schuff, and Zhewei Zhang. Why aren't the stars aligned? an analysis of online review content and star ratings. In *2014 47th Hawaii International conference on system sciences*, pages 3139–3147. IEEE, 2014.

Hugo Gonçalo Oliveira, Inês Coelho, and Paulo Gomes. Exploiting portuguese lexical knowledge bases for answering open domain cloze questions automatically. In *LREC*, pages 4202–4209, 2014.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362, 2022.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. A survey of text representation and embedding techniques in nlp. *IEEE Access*, 2023.

Rafael Peres, Diego Esteves, and Gaurav Maheshwari. Bidirectional lstm with a context input window for named entity recognition in tweets. In *Proceedings of the knowledge capture conference*, pages 1–4, 2017.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266*, 2022.

Ramon Pires, Hugo Abonizio, Thales Rogério, and Rodrigo Nogueira. Sabi\'a: Portuguese large language models. *arXiv preprint arXiv:2304.07880*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Livy Real, Marcio Oshiro, and Alexandre Mafra. B2w-reviews01-an open product reviews corpus. In *the Proceedings of the XII Symposium in Information and Human Language Technology*, pages 200–208, 2019.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

João Rodrigues, António Branco, Steven Neale, and João Silva. Lx-dsemvectors: Distributional semantics models for portuguese. In *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings 12*, pages 259–270. Springer, 2016.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. Advancing neural encoding of portuguese with transformer albertina pt. *arXiv preprint arXiv:2305.06721*, 2023.

Ruan Chaves Rodrigues, Jéssica Rodrigues, Pedro Vitor Quinta de Castro, Nádia Felix Felipe da Silva, and Anderson Soares. Portuguese language models and word embeddings: evaluating on semantic similarity tasks. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 239–248. Springer, 2020.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243.

Umitcan Sahin, Oguzhan Ozcelik, Izzet Emre Kucukkaya, and Cagri Toraman. ARC-NLP at CASE 2022 task 1: Ensemble learning for multilingual protest event detection. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 175–183, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.case-1.25. URL https://aclanthology.org/2022.case-1.25.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. Harem: An advanced ner evaluation contest for portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*, 2006.

Joaquim Santos, Bernardo Consoli, and Renata Vieira. Word embedding evaluation in downstream tasks and semantic analogies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4828–4834, 2020.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448. IEEE, 2019.

Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training, 2019.

Alberto Manuel Brandão Simões. Desafio de identificação de personagens. *Linguamática*, 15(1):iii–ix, 2023.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 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, 2020.

P. Szymański and T. Kajdanowicz. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*, February 2017.

Maksim A Terpilowski. scikit-posthocs: Pairwise multiple comparison tests in python. *Journal of Open Source Software*, 4(36):1169, 2019.

Ankit Vaidya and Aditya Kane. Two-stage pipeline for multilingual dialect detection. *arXiv preprint arXiv:2303.03487*, 2023.

Raphael Vallat. Pingouin: statistics in python. *J. Open Source Softw.*, 3(31):1026, 2018.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 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, 2022.

David Vilares, Marcos Garcia, and Carlos Gómez-Rodríguez. Bertinho: Galician bert representations. *arXiv preprint arXiv:2103.13799*, 2021.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1686`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015.

Arkaitz Zubiaga. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984, 2018.