Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family

Rodrigo Santos[†], **João Rodrigues**[†], **Luís Gomes**[†], **João Silva**[†], **António Branco**[†], **Henrique Lopes Cardoso**[‡], **Tomás Freitas Osório**[‡], **Bernardo Leite**[‡]

[†]University of Lisbon
NLX - Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal
{rsdsantos, jarodrigues, luis.gomes, antonio.branco}@fc.ul.pt

[‡]University of Porto
Faculty of Engineering, Department of Informatics Engineering
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
hlc@fe.up.pt, tomas.s.osorio@gmail.com, bernardo.leite@fe.up.pt

Abstract

To foster the neural encoding of Portuguese, this paper contributes foundation encoder models that represent an expansion of the still very scarce ecosystem of large language models specifically developed for this language that are fully open, in the sense that they are open source and openly distributed for free under an open license for any purpose, thus including research and commercial usages. Like most languages other than English, Portuguese is low-resourced in terms of these foundational language resources, there being the inaugural 900 million parameter Albertina and 335 million Bertimbau. Taking this couple of models as an inaugural set, we present the extension of the ecosystem of state-of-the-art open encoders for Portuguese with a larger, top performance-driven model with 1.5 billion parameters, and a smaller, efficiency-driven model with 100 million parameters. While achieving this primary goal, further results that are relevant for this ecosystem were obtained as well, namely new datasets for Portuguese based on the SuperGLUE benchmark, which we also distribute openly.

Keywords: Large language model, foundation model, encoder, Portuguese, open-source

1. Introduction

The present paper contributes foundation models that represent the development and the populating of the still very scarce ecosystem of fully open large language models of the encoder family of Transformers specifically developed for the Portuguese language, that is models that are open source and openly distributed with for free with an open license.

Since their appearance in (Vaswani et al., 2017) and given their superior performance vis a vis their viable alternatives, neural language models based on the Transformer architecture became the mainstream approach for virtually any natural language processing task (Brown et al., 2020; Raffel et al., 2020; He et al., 2021). Transformers were proposed in an encoder-decoder setup (Raffel et al., 2020), but encoder-only and decoder-only setups have also been shown highly competitive by subsequent research (Devlin et al., 2019; He et al., 2021; Brown et al., 2020).

Despite the outstanding visibility that the Transformer-based decoder models have deservedly garnered, especially with the availability of ChatGPT for the general public, the models of the encoder family have not lost their traction as they have maintained a competitive performance in non-generative tasks, especially in those tasks primarily related to classification (He et al., 2021; Zhong et al., 2022).[1]

The largest and more powerful foundation models have been developed for English — (He et al., 2021; Touvron et al., 2023) among many others —, which is the language that, among the more than 7 000 idioms on the planet, is by a very large margin the one whose research is better funded, better technologically prepared for the digital age and for which more language resources have been developed (Rehm and Way, 2023).

Additionally, multilingual models have also been developed, whose training is done over datasets that extend its majority of English data with proportionally much smaller data portions from a few other languages (Devlin et al., 2019; Chowdhery et al., 2022; Scao et al., 2022). Leveraged by the sheer volume of data thus made available, these models have shown competitive performance in handling tasks in the languages, other than English, whose data portions are a minority in their training set (Wu and Dredze, 2019).

On par with these results and their relevance for

[1]At the time of writing, as a way of confirmation of this remark, the top performing model in the SuperGLUE benchmark (https://super.gluebenchmark.com/leaderboard) is an encoder, namely the Vega v2 model (Zhong et al., 2022).

SIGUL2024 Workshop, pages 105–114
21-22 May, 2024.

some multilingual natural language tasks, especially machine translation, other approaches have been explored, namely with the continuation of the pre-training of multilingual or plain English models with data from a specific language. Reported results seem to converge in indicating that when their continued training is appropriately setup, the performance of the resulting models on language-specific tasks shows important improvements over a possible baseline model whose training was performed from scratch with the same (comparatively small) amount of language-specific data (Kim et al., 2021; Pires et al., 2023; Rodrigues et al., 2023).

Adopting this latter approach and adding to the previous work on the neural encoding of Portuguese (Rodrigues et al., 2023; Souza et al., 2020), the present paper puts forward further models for this language that expand its ecosystem of open encoders. These encoders cumulatively comply with all the features of being open source, publicly available for free, and distributed under a most permissive license (including for research and for commercial purposes). Furthermore, they are available for two variants of Portuguese: European Portuguese, spoken in Portugal (PTPT), and American Portuguese, spoken in Brazil (PTBR).

Taking as reference the existing state-of-the-art 900 million parameter encoder Albertina (Rodrigues et al., 2023), which complies with all the above requirements, in this paper we present the extension of the ecosystem of open encoders for Portuguese with a larger, top performance-driven encoder model with 1.5 billion parameters, Albertina 1.5B PT, and a smaller, efficiency-driven encoder model with 100 million parameters, Albertina 100M PT. These models are distributed from `https://huggingface.co/PORTULAN`.

While achieving these central goals, further results that are relevant for this ecosystem were obtained as well: new datasets for Portuguese based on the trusted GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks, which are distributed openly; and state-of-the-art performance for Portuguese in various natural language processing tasks in these benchmarks.

The remainder of this paper is structured as follows: the next Section 2 discusses related work. In Section 3 the data used in the creation of the various models is presented; the encoder models created in this study are described in Section 4; Section 5 presents the evaluation results; and Section 6 closes the paper with concluding remarks.

2. Related Work

The advent of the Transformer architecture (Vaswani et al., 2017) represents a revolutionary milestone in the field of Natural Language Processing. With its attention mechanisms, the Transformer enabled the efficient modeling of contextual information in text, paving the way for the development of powerful models.

The success of this architecture led to the emergence of various encoder models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021), which set new standards for language comprehension tasks. Nevertheless, they cater exclusively for the English language.

To address linguistic diversity, multilingual encoder models emerged as a promising solution. Notable examples include mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020), among others, which support multiple languages and seek to bridge language barriers.

In contrast, a few encoder models that cater for specific languages have also been introduced. For instance, ERNIE (Sun et al., 2021) for Chinese, CamemBERT (Martin et al., 2020) for French, and MarIA (Gutiérrez-Fandiño et al., 2022) for Spanish, among others. These have demonstrated the importance of language-tailored models in capturing language-specific nuances, which multilingual models cannot so easily ensure (Papadimitriou et al., 2023).

Concerning Portuguese, previous encoder models such as the 900 million parameter Albertina (Rodrigues et al., 2023) and the 335 million parameter BERTimbau (Souza et al., 2020) have made significant contributions. With BERTimbau covering PTBR, and Albertina covering both PTPT and PTBR variants, these models have not only bolstered the Portuguese NLP ecosystem but have also set the path for the development of more advanced language models tailored to the Portuguese language.

In this paper, we aim at adding to this existing work by contributing further encoder models with further dimensions, also covering both the European PTPT and the American PTBR variants of Portuguese.

3. Data

In this section, we present the data used for the training and testing of our encoder models.

In both their variants, PTBR and PTPT, for our smaller, 100 million parameter model, we resort to the Portuguese subset of the OSCAR dataset (Abadji et al., 2022). And for our larger, 1.5 billion parameter model, we resort to the Portuguese subset of the CulturaX dataset (Nguyen et al., 2023). Additionally, for the models handling the PTPT variants, the dataset we used included also the monolingual corpora DCEP, ParlamentoPT and Europarl

dataset	exs (M)	words (B)
Albertina 100M PTPT	10.2	2.4
Albertina 100M PTBR	4.1	2.7
Albertina 1.5B PTPT	16.1	4.3
Albertina 1.5B PTBR	87.9	36.2

Table 1: Size of datasets used for training, in millions of examples (exs) and in billions of words

(Hajlaoui et al., 2014; Koehn, 2005; Rodrigues et al., 2023).

These corpora and their curation are described in detail below in the next Subsection, and their sizes are summarized in Table 1.

3.1. Training Data

While both multilingual datasets, OSCAR and CulturaX, distribute their Portuguese subsets separately, they do not provide further separation between European Portuguese and American Portuguese within these subsets. To separate the texts in one variant from the texts in the other, we use the source URLs provided with every data entry and filter by top-level domain. We only keep entries with the ".br" top-level domain, and add them to the PTBR subset, and with the ".pt" top-level domain, for the PTPT subset.

From these datasets, data entries of domains whose content should not be redistributed were removed, in order to limit the possibility of content reproduction by the models or by future derivatives that will resort to these datasets.

OSCAR Corpus The project promoting the OSCAR corpus is an open source project which distributes multilingual datasets for machine learning and artificial intelligence applications (Abadji et al., 2022).

The OSCAR subset for Portuguese we use is based on November/December 2022 version of Common Crawl, which is an automatic crawl from the web. Despite being a crawl, the final dataset is of relatively good quality due the filtering performed on the corpus by its authors. As can be seen in Table 2, we end up with subsets of OSCAR for the two Portuguese variants that have a not too distinct number of examples and words.

CulturaX Corpus CulturaX is a multilingual corpus, freely available for research and AI development (Nguyen et al., 2023), created by combining and extensively cleaning two other large datasets, mC4 (Xue et al., 2021) and OSCAR.

The CulturaX subset for PTBR is an order of magnitude larger than for PTPT, as depicted in Table 2, both in examples and words. This does not present itself as a problem since we aim to develop the best model possible for each variant.

dataset	examples (M)	words (M)
OSCAR ptbr	4.1	2,728
OSCAR ptpt	3.0	1,976
CulturaX ptbr	87.9	36,201
CulturaX ptpt	8.9	3,896
DCEP	2.5	76
ParlamentoPT	2.9	289
Europarl	1.8	49

Table 2: Number of examples and words for each dataset for training

Other Corpora In addition to the above language resources, for the European Portuguese versions we also include in our training set: (i) the Portuguese portion of DCEP (Hajlaoui et al., 2014), a Digital Corpus of the European Parliament; (ii) the Portuguese portion of Europarl (Koehn, 2005), the European Parliament Proceedings Parallel Corpus; and (iii) ParlamentoPT (Rodrigues et al., 2023), a corpus of transcriptions of the debates in the Portuguese Parliament.

These corpora are based on human transcriptions of parliamentary debates and can be assumed to be of very high quality, despite their limited domain. They provide a good complement to OSCAR and CulturaX.

Finally, we apply further quality filtering to all corpora—except to CulturaX, since it already has a good quality filtering step—, through the use of the Bloom pre-processing pipeline (Laurençon et al., 2022).

Table 2 presents statistics for all the corpora used in this work; all these numbers are calculated right before training the model, i.e. after splitting between variants and applying all types of additional content filtering.

3.2. Testing Data

The performance of encoder models are typically evaluated by testing them in downstream tasks. For the Portuguese language, both variants, there is however a lack of such datasets, either in quality or in quantity, to appropriately evaluate an encoder models. The only dataset created from scratch in (American) Portuguese, that we could find, is the ASSIN 2 dataset (Real et al., 2020) that was used to evaluate BERTimbau.

To cope with this hindrance, we contribute new test datasets for Portuguese based on the GLUE

(Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks.

We obtain these datasets through machine translation from English using DeepL,[2] which allows translation either to PTPT or to PTBR, and to our knowledge is the only online service that translates to both of these variants. DeepL is regarded as one of the best machine translation services available online.[3]

The exception to this translation process, concerns the PTBR portion of GLUE, which we took from PLUE (Gomes, 2020), to avoid redoing valid work already present in the literature and openly distributed.

ASSIN 2 tasks The ASSIN 2 dataset contains two tasks: (i) RTE, for recognizing textual entailment, and (ii) STS, for semanting textual similarity.

GLUE tasks From GLUE we chose four tasks: two similarity tasks, (i) MRPC, for detecting whether two sentences are paraphrases of each other, and (ii) STS-B, for semantic textual similarity; and two inference tasks, (iii) RTE, for recognizing textual entailment, and (iv) WNLI, for coreference and natural language inference.

SuperGLUE tasks As for SuperGlue, we also chose four tasks: two QA tasks, (i) MultiRC, for detecting whether an answer to a question about a paragraph is correct or not, and (ii) BoolQ, for answering *yes* or *no* to a question about a passage; one reasoning task, (ii) COPA, given a premise sentence and two possible choices, the system must determine either the cause or effect of the premise from two possible choices; and one inference task with three labels, (iv) CB, for predicting how much the text commits to the clause.

4. Models

This section describes the training of the models contributed in this paper.

4.1. The starting models

We use DeBERTa (He et al., 2021) as a starting point from which to continue the pre-training of our models over Portuguese data. This is an encoder that incorporates a new attention mechanism, making it particularly effective for a wide range of natural language processing tasks. DeBERTa's architecture disentangles attention patterns, improving its ability to capture relationships between words and phrases in a text.

With its different model sizes, including the compact DeBERTa-Base with 100 million parameters, the DeBERTa-XLarge with 900 million parameters, and the high-capacity DeBERTa-XXLarge with 1.5 billion parameters, it caters for various NLP requirements.

The only encoder for both variants PTP and PTBR variants of Portuguese, the existing 900 million parameter model Albertina, was obtained by continuing the pre-training of DeBERTa-XLarge with Portuguese (Rodrigues et al., 2023).

With the same goal in mind, we start from the DeBERTa-Base to construct our Albertina 100M PT models, and from the DeBERTa-XXLarge, for our Albertina 1.5B PT models.

4.2. The Albertina 100M PT Foundation Model

The two smaller models, Albertina 100M PTPT and Albertina 100M PTBR, are constructed upon the DeBERTa Base V1 model, comprising 100 million parameters.

The models were trained on a a2-megagpu-16gb Google Cloud A2 node equipped with 16 GPUs, 96 vCPUs, and 1.360 GB of RAM, and their training took approximately one day of compute. This configuration resulted in a batch size of 3072 samples, with 192 samples allocated per GPU, when trying to fill the whole memory available.

We used the original DeBERTa tokenizer for both models, implementing a 128-token sequence truncation and dynamic padding. The training was performed under a learning rate of 1e-5, with linear decay and 10k warm-up steps, determined after a few exploratory trials. The PTPT model underwent 200 training epochs, while the PTBR model underwent 150, accumulating roughly 180k training steps in each case.

4.3. The Albertina 1.5B PT Foundation Model

As for the larger models, Albertina 1.5B PTPT and PTBR, we developed them upon the DeBERTa XXLarge V2 encoder, comprising 1.5 billion parameters.

Similarly to the smaller models, the two Albertina 1.5B PTmodels were trained on a a2-megagpu-16gb Google Cloud A2 node.

We resorted to the original DeBERTa V2 tokenizer for both models, implementing a 128-token sequence truncation and dynamic padding for 250k steps, a 256-token sequence-truncation for 80k steps and finally a 512-token sequence-truncation for 60k steps. These steps correspond to the equivalent setup of 48 hours on a2-megagpu-16gb

[2] https://www.deepl.com/

[3] The construction is thoroughly presented in (Osório et al., submited)

Google Cloud A2 node for the 128-token input sequences, 24 hours of computation for the 256-token input sequences and 24 hours of computation for the 512-token input sequences.

We applied a learning rate of 1e-5, with linear decay and 10k warm-up steps, determined after a few exploratory trials

5. Evaluation and Discussion

This section presents and discusses the evaluation of our models, introduced just above in Section 4, with respect to the downstream tasks, introduced in Section 3.2, after their fine-tuning on these tasks.

Additionally, for the sake of a thorough comparative evaluation of these models, this section also presents the results of fine-tuning and evaluating in the same downstream tasks, the pre-existing models in the ecosystem of encoders for Portuguese, namely the 900 million parameter Albertina and the 335 million parameter BERTimbau We also evaluate with the two DeBERTa baseline models, with 100 million and 1.5 billion parameter, trained mostly with English data, which we did not continue the training on further Portuguese data.

The compilation of all these results are in Table 4, for the model versions concerning the PTBR variant, and Table 5, for the PTPT variant.

5.1. Fine-tuning

Each model under evaluation was fine-tuned on each of the eight downstream tasks obtained from GLUE and SuperGLUE and introduced in Section 3.2.[4] In order to proceed with hyper-parameter optimization, the following hyper-parameter values were chosen for our grid-search:

- Epochs: 5
- Batch size: 4
- Learning rate: $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-6}\}$
- Learning rate scheduler type: linear
- Warm up ratio: 0.1
- Adam epsilon: 1×10^{-6}
- Weight decay: 0.01
- Dropout: $\{0, 0.1\}$
- BF16: $\{0, 1\}$

A hyper-parameter grid search was performed for each pre-trained model/task combination, resulting in a total of **4104** fine-tuned and evaluated models. This number results from 12 combinations of hyper-parameter values (3 learning rates $\times$ 2 dropout values $\times$ 2 BF16 values), times the number of tasks (10 for PT-BR and 8 for PT-PT), times the number of evaluated pre-trained models[5] (7 for PT-BR and 6 for PT-PT), times 3 random seeds.

As presented in Section 3, the GLUE and SuperGLUE evaluation datasets were translated into both Portuguese variants from their English originals.

It is noteworthy that the test sets from from the GLUE and SuperGLUE datasets are not distributed with ground labels, as evaluation is setup to proceed by submitting online the data to be evaluated. Given that the number of such online submissions per month for each user is highly limited and very small, and given the very large number of models and tasks and thus of evaluation runs we needed to cope with, it was not practically viable to resort to such online evaluation service. As a consequence, to proceed with our very large experimental space, we adopted the same methodology as we did for the 900 million parameter Albertina (Rodrigues et al., 2023): we used the validation partitions of the downstream datasets for testing; and for training, we randomly split the partition that is originally distributed for training into 90% that we used for actual training and into the remaining 10% that we used for development and validation purposes.

After acquiring the best hyper-parameter values on the data that were set aside for development purposes and by using such hyper-parameters, the performance scores were obtained by testing on the subsets that were left for evaluation, which are displayed in Tables 3, 4 and 5. The values presented are the average scores of 3 runs with different random seeds.

5.2. Albertina 1.5B PT Fine-tuned

Since most tasks have input sizes closer to 256 than to 512, we evaluated two variants of the Albertina 1.5B PT model: the models with suffix S (short) in Tables 4 and 5 are fine-tuned from checkpoints after pre-training with sequences of 256 tokens; while the models with suffix L (long) are fine-tuned from the final checkpoints, i.e. after pre-training with sequences of 512 tokens.

In almost all tasks and for both language variants, our largest model, with 1.5 billion parameters, shows the best performance scores, and in the few cases where that is not the case, it competitively come close to the best scoring model.

It is of note that among the downstream tasks, WNLI appears somehow as an outlier as the per-

[4] The exception were the 100M DeBERTa models (DeBERTa-base and both versions of Albertina 100M PT), which were not evaluated on the COPA task because the Hugging Face head for multiple choice does not support DeBERTa v1 models.

[5] The 100M parameter models could not be evaluated in the COPA task for lack of support for these models in the HuggingFace head implementation for this task.

model	ASSIN2 RTE	ASSIN2 STS
Albertina 1.5B PTBR L	**0.9153**	0.8647
Albertina 1.5B PTBR S	0.9109	**0.8688**
Albertina 900M PTBR	0.9130	0.8676
BERTimbau (335M)	0.8913	0.8531
Albertina 100M PTBR	0.8747	0.8269
DeBERTa 1.5B EN	0.8803	0.8356
DeBERTa 100M EN	0.8369	0.7760

Table 3: Evaluation scores for **PTBR** on the ASSIN2 native American Portuguese dataset. Performance on RTE is measured with accuracy and on STS with Pearson

model	GLUE RTE	GLUE WNLI	GLUE MRPC	GLUE STS-B	SuperGLUE COPA	SuperGLUE CB	SuperGLUE MultiRC	SuperGLUE BoolQ
Albertina 1.5B PTBR L	**0.8676**	0.4742	0.8622	**0.9007**	0.7767	0.6372	**0.7667**	**0.8654**
Albertina 1.5B PTBR S	0.8123	0.4225	0.8638	0.8968	**0.8533**	**0.6884**	0.6799	0.8509
Albertina 900M PTBR	0.7545	0.4601	**0.9071**	0.8910	0.7767	0.5799	0.6731	0.8385
BERTimbau (335M)	0.6446	0.5634	0.8873	0.8842	0.6933	0.5438	0.6787	0.7783
Albertina 100M PTBR	0.6582	**0.5634**	0.8149	0.8489	n.a.	0.4771	0.6469	0.7537
DeBERTa 1.5B EN	0.7810	0.4789	0.8555	0.8600	0.4733	0.4648	0.6738	0.8315
DeBERTa 100M EN	0.5716	0.5587	0.8060	0.8266	n.a.	0.4739	0.6391	0.6838

Table 4: Evaluation scores for **PTBR**. Performance on RTE, WNLI, BoolQ and COPA is measured with accuracy, on MRPC, MultiRC and CB with F1, and on STS-B with Pearson

formance level of the different models on it is not aligned with their performance level in the other tasks. This has been already observed also with Albertina 900 M (Rodrigues et al., 2023), which attributed this to the very small size of the WNLI dataset.

In its overall performance, this largest model surpasses the previously best model Albertina 900M in this ecosystem, and offers thus the state-of-the-art performance in most tasks for Portuguese by an open encoder.

5.3. Albertina 100M PT Fine-tuned

With 100 million parameters, our Albertina 100M PT model is the smallest in this ecosystem of open encoders for Portuguese. Yet, it has very good performance taking into account its reduced size.

Taking WNLI aside, Albertina 100M PT matches or surpasses its base model (DeBERTa 100M) in all 16 tasks, except in CB for PTPT.

On the other hand, our Albertina 100M PTBR is very competitive with respect to the BERTimbau model, whose 335 million parameters are more than the triple of its size. It surpasses BERTimbau's performance in GLUE's RTE, and supports a very competitive second position in most of the other tasks. Likely, this is the consequence of BERTimbau having BERT (Devlin et al., 2019) as its base model, while Albertina 100M PTis based in the more advanced DeBERTa (He et al., 2021).

5.4. Discussion

The larger the better Taking a broad view of the results in Tables 4 and 5, overall and as expected, the larger the Albertina model the better is its performance in downstream tasks.

In this respect, and taking aside WNLI, already commented on above, the exception to this trend is MRPC. In this task, the 1.5B Albertina models are outperformed by the smaller 900M Albertinas. Although we don't have a compelling explanation for this, it appears that the 900M parameter network may provide the optimal expressive power for learning this particular task and dataset, across the various model sizes under evaluation.

The more monolingual the better When compared to their respective DeBERTa baseline counterparts, our newly contributed models, Albertina 1.5B PTand Albertina 100M PT, present superior performance in general.

This adds to the empirical evidence in the literature, commented in Section 2, for the importance of continuing the pre-training of models with monolingual data for the language of interest, even if they started multilingual or were initially developed for another language. If appropriately prepared, the resulting models typically represent a better solution

model	GLUE				SuperGLUE			
	RTE	WNLI	MRPC	STS-B	COPA	CB	MultiRC	BoolQ
Albertina 1.5B PTPT L	**0.8809**	0.4742	0.8457	**0.9034**	**0.8433**	**0.7840**	**0.7688**	**0.8602**
Albertina 1.5B PTPT S	0.8809	0.5493	0.8752	0.8795	0.8400	0.5832	0.6791	0.8496
Albertina 900M PTBR	0.8339	0.4225	**0.9171**	0.8801	0.7033	0.6018	0.6728	0.8224
Albertina 100M PTPT	0.6919	0.4742	0.8047	0.8590	n.a.	0.4529	0.6481	0.7578
DeBERTa 1.5B EN	0.8147	0.4554	0.8696	0.8557	0.5167	0.4901	0.6687	0.8347
DeBERTa 100M EN	0.6029	**0.5634**	0.7802	0.8320	n.a.	0.4698	0.6368	0.6829

Table 5: Evaluation scores for **PTPT**. Performance on RTE, WNLI, BoolQ and COPA is measured with accuracy, on MRPC, MultiRC and CB with F1, and on STS-B with Pearson

for that language.

Concerning the largest model Abertina 1.5B, and taking aside the WNLI outlier, it always improves over its baseline model.

As for our smaller model Albertina 100M, the exception to this trend appears once again in WNLI, for PTPT, and CB, by a small margin, also for PTPT.

The more advanced the base model the better Comparing the new Albertina 100M PT and Albertina 1.5B PT models to the previously existing models, it is clear that the larger models offer improvements over smaller models as noted above.

However, it is important also to note that the difference between the performance scores of Albertina 100M PTBR and of the 335M BERTimbau is rather small, which seems to suggest that the improvements in DeBERTa, on which our Albertina 100M PTis based, over BERT, which used as a base model by BERTimbau, have allowed for more efficient parameter utilization and improved performance in general.

The more language variants the better For the same task and the same model dimension, the models for the European PTPT and American PTBR variants of Portuguese show different performance scores. While in general not representing a wide gap, these differences exist, as expected.

These differences should be attributed, for instance, to the possible different quality of the translations produced for the English datasets, depending on the Portuguese variant, and also attributed in some cases to the different sizes of the training corpora, etc. For instance, the training of the 1.5 billion model for PTBR was based on a 36.2 billion token dataset, while the same size model for PTPT resorted to a much smaller, 4.3 billion token corpus, as indicated in Table 1.

From the three models with two versions, i.e. one version per variant, namely, the Albertina 100M, 900M and 1.5B models, it is the 900M one than may permit a more insightful comparison among its two variants given the conditions of their training were closer to each other, with a 2.7M and a 2.2M token training dataset for PTBR and PTPT, respectively (Rodrigues et al., 2023).

Thus looking to the experimental results we obtained for the two Albertina 900M versions, PTBR and PTPT, across the Tables 4 and 5, one finds deltas, for instance, of 0.079 (accuracy) in RTE, 0.073 (F1) in COPA, or 0.022 (accuracy) in CB. This is in line with the same lessons drawn in (Rodrigues et al., 2023), and it is confirming its results. It is thus relevant to keep the two variants of Portuguese addressed by different model versions if possible.

6. Conclusions

The results reported in the present paper demonstrate that the models hereby contributed represent valuable advances for the ecosystem of fully open large language models of Portuguese.

With its 1.5 billion parameters, Albertina 1.5B PT becomes the largest open encoder specifically developed for this language, and the one that better support state of the art performance in downstream tasks.

With its 100 million parameter, Albertina 100M PT becomes, in turn, the smallest, appropriately curated and documented, open encoder of this ecosystem, and thus the one that ensures an encoding solution for this language that favours efficiency and is available to run in limited hardware.

It is also worth noting that the advancements contributed in this paper for both American and European variants of Portuguese cater for the linguistic diversity in this language, ensuring their relevance and applicability to a broad user base.

In conclusion, this paper presents a significant contribution to the field of language technology for Portuguese by introducing state-of-the-art large language models that serve the technological preparation of this language. The models are not only technically robust but also fully open, in the sense that are open source, openly distributed for free under an open license for both research and commercial

purposes. They are adaptable for various applications, thus facilitating innovation and progress in the field.

These models can be obtained from https://huggingface.co/PORTULAN.

Future work will include further expanding and updating this ecosystem of fully open encoders for Portuguese with other model dimensions, other language variants and other design features.

Acknowledgements

This research was partially supported by: PORTULAN CLARIN — Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT (PINFRA/22117/2016); ACCELERAT.AI - Multilingual Intelligent Contact Centers, funded by IAPMEI (C625734525-00462629); ALBERTINA - Foundation Encoder Model for Portuguese and AI, funded by FCT (CPCA-IAC/AV/478394/2022); and LIACC - Artificial Intelligence and Computer Science Laboratory (FCT/UID/CEC/0027/2020).

7. Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 4344–4355.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

J. R. S. Gomes. 2020. Plue: Portuguese language understanding evaluation. https://github.com/ju-resplande/PLUE.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, pages 39–60.

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, and Daniel Varga. 2014. DCEP—Digital corpus of the European Parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 2021. What changes can large-scale language models bring? Intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: papers*, pages 79–86.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The BigScience ROOTS corpus: A 1.6 TB composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Tomás Freitas Osório, Bernardo Leite, Henrique Lopes Cardoso, Luís Gomes, João Rodrigues, Rodrigo Santos, and António Branco. submited. Extraglue datasets and models: Kickstarting a benchmark for the neural processing of portuguese.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 143–146. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21:5485–5551.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 406–412. Springer.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of Portuguese with transformer Albertina PT-*. In *Progress in Artificial Intelligence (EPIA 2023)*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018.

GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on SuperGLUE. *arXiv preprint arXiv:2212.01853*.

8. Language Resource References

Fábio Souza and Rodrigo Nogueira and Roberto Lotufo. 2020. *BERTimbau Large*. Hugging Face.

J. R. S. Gomes. 2020. *PLUE: Portuguese Language Understanding Evaluation*. Hugging Face.

Hajlaoui Najeh, Kolovratnik David, Vaeyrynen Jaakko, Steinberger Ralf, and Varga Dániel. 2012. *DCEP: Digital Corpus of the European Parliament*. European Parliament - DG TRAD. European Parliament - DG TRAD, ISLRN 823-807-024-162-2.

João Rodrigues and Luís Gomes and João Silva and António Branco and Rodrigo Santos and Henrique Lopes Cardoso and Tomás Osório. 2023a. *Albertina PT-BR*. PORTULAN CLARIN. distributed via PORTULAN CLARIN. PID https://hdl.handle.net/21.11129/0000-000F-FF43-7.

João Rodrigues and Luís Gomes and João Silva and António Branco and Rodrigo Santos and Henrique Lopes Cardoso and Tomás Osório. 2023b. *Albertina PT-PT*. PORTULAN CLARIN. distributed via PORTULAN CLARIN. PID https://hdl.handle.net/21.11129/0000-000F-FF42-8.

Julien Abadji and Pedro Ortiz Suarez and Laurent Romary and Benoît Sagot. 2023. *OSCAR 23.01 – Open Source Project on Multilingual Resources for Machine Learning*. the OSCAR project.

Pengcheng He and Xiaodong Liu and Jianfeng Gao and Weizhu Chen. 2023. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. Microsoft.

Philipp Koehn. 2012. *European Parliament Proceedings Parallel Corpus (v7)*. EuroMatrixPlus project.

Real, Livy and Fonseca, Erick and Gonçalo Oliveira, Hugo. 2020. *ASSIN 2 (The ASSIN 2 Shared Task: A Quick Overview)*. Hugging Face.

Thuat Nguyen and Chien Van Nguyen and Viet Dac Lai and Hieu Man and Nghia Trung Ngo and Franck Dernoncourt and Ryan A. Rossi and Thien Huu Nguyen. 2023. *CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages*. Hugging Face.

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

Wang, Alex and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel. 2018. *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. Hugging Face.