An evaluation of Portuguese language models' adaptation to African Portuguese varieties

Diego Alves
Saarland University / Saarbrücken, Germany
diego.alves@uni-saarland.de

Abstract

In this study, we conduct a comparative evaluation of two state-of-the-art language models, Albertina PT-PT and Albertina PT-BR, which are trained on European Portuguese and Brazilian Portuguese, respectively. Our aim is to assess their suitability for African varieties of Portuguese. To evaluate their performance, we create two test sets for each variety, encompassing both spoken and written language. We measure the percentage of sentences in which one model outperforms the other in terms of perplexity. This evaluation seeks to ascertain whether one model shows more adaptability to the African varieties of Portuguese. Our findings reveal that Albertina PT-PT consistently outperforms Albertina PT-BR in scenarios involving spoken language corpora. However, in written registers, the advantage of Albertina PT-PT is less pronounced for the Portuguese varieties of Guinea-Bissau, Mozambique, and São Tomé and Principe. These insights contribute to our understanding of the adaptability of existing language models to African Portuguese varieties and emphasize the need for specialized models to address the unique linguistic nuances of this region.

1 Introduction

In recent years, language modeling (LM) has become one of the major strategies for advancing Natural Language Processing (NLP) showing strong capabilities in improving scores in a large variety of tasks. Basically, its aim is to model the generative likelihood of word sequences, in order to predict the probabilities of future (or missing) tokens (Zhao et al., 2023).

As was the case in other NLP fields, the development of pre-trained language models primarily focused on the English language (Chowdhery et al., 2022). However, this technology has been deployed to other languages, especially major ones, with the development of language-specific language models or multilingual ones such as Multilingual BERT (Pires et al., 2019), XLM-R (Conneau et al., 2019), mBART (Liu et al., 2020), mT5 (Xue et al., 2020), and BLOOM (Scao et al., 2022).

Regarding Portuguese, in the "Report on the Portuguese Language" of the European Language Equality consortium (Branco et al., 2022), the conclusion is that there is a severe lack of freely available, last-generation large language models. The situation is even more critical for African varieties of Portuguese as the existing Portuguese language models have been trained only or mostly with European and Brazilian Portuguese corpora.

Portuguese is spoken in 6 African countries: Angola, Cape Verde, Equatorial Guinea, Guinea-Bissau, Mozambique, and São Tomé and Príncipe. In these countries, Portuguese is not the main native language but in Angola, Mozambique, and Cape Verde, it is spoken at least by 40% of the population (Eberhard et al., 2023).

As of now, due to the lack of pre-trained language models specifically aimed at African varieties of Portuguese, researchers dealing with the development of NLP tools for these varieties do not have another choice rather to use one of the multilingual or language-specific models trained on European and/or Brazilian Portuguese.

Thus, the aim of this article is to present a comparative study regarding the processing of different African varieties of the Portuguese language with state-of-the-art Portuguese language models.

With this intention, we analyse how well a language model trained with European Portuguese texts performs when processing different African varieties of Portuguese in comparison to a similar model (in terms of training parameters) trained with Brazilian Portuguese texts.

The remainder of this paper is structured as follows. First, we present the related work, then, Section 3 describes the methodology regarding the corpora acquisition and the perplexity measures. In

Section 4, we present the obtained results, followed by a discussion in Section 5. Finally, Section 6 is dedicated to the main conclusions and perspectives for future work.

2 Related work

It has been shown that language-specific models tend to be better for a large variety of NLP tasks when compared to multilingual ones (e.g., Devlin et al. (2018); Virtanen et al. (2019); De Vries et al. (2019); Martin et al. (2019)). Multilingual language models are a useful solution in cases where a specific language model does not exist due to a lack of available data or data processing resources.

Regarding the Portuguese language, the most used language models concerning general tasks are multilingual: XML-R (Conneau et al., 2019) and Multilingual BERT (mBERT) (Pires et al., 2019).

Among the publicly available models for Portuguese, BERTabaporu (da Costa et al., 2023) is a BERT-based encoder trained on Brazilian Portuguese Twitter data. It was built using a collection of 238 million tweets written by over 100,000 unique Twitter users (over 2.9 billion tokens in total).

However, the most popular encoder for PT-BR is BERTimbau (Souza et al., 2020) as it covers a larger variety of genres. It is available in two model sizes (110 million parameters and 330 million parameters) and both variants were trained with the brWaC corpus (Wagner Filho et al., 2018) having a BERT-based model as a starting point. These models outperform mBERT in many NLP tasks as shown by Souza et al. (2020).

The lack of publicly available European Portuguese language models and the work developed regarding BERTimbau inspired the creation of the Albertina PT transformers (Rodrigues et al., 2023) covering two varieties of Portuguese: European Portuguese from Portugal (PT-PT) and American Portuguese from Brazil (PT-BR). These models were developed using DeBERTa as a starting point. For Albertina PT-PT, a specific training corpus was gathered, and regarding Albertina PT-BR, brWaC was used (same as BERTimbau). The evaluation provided by the authors showed that Albertina PT-BR outperforms BERTimbau in several tasks and Albertina PT-PT provides interesting results for the European variant of the Portuguese language.

If some work has been developed regarding Brazilian and European Portuguese (although incipient when compared to English), regarding African varieties of Portuguese have been completely neglected. The development of large language models for African languages has focused on indigenous languages. It is the case of AfriBERTa (Ogueji et al., 2021) and Afro-XLMR-base (Alabi et al., 2022). Only SERENGETI model (Adebara et al., 2022) includes Creole Portuguese in its set of languages.

Therefore, due to the lack of evaluation of Portuguese language models for African varieties of Portuguese, we decided to conduct a comparative analysis of Albertina models to check which version (PT-PT or PT-BR) is more adapted to be used in NLP tasks regarding African varieties of Portuguese. Albertina models have been chosen as they can be considered state-of-the-art for Portuguese and because both PT-PT and PT-BR are comparable in terms of parameters (although diverse in terms of training data).

Our objective is to contribute to the understanding of how language models perform regarding different varieties of Portuguese (until now ignored) and to inspire further variety-specific developments.

We decided to use perplexity measures as it is the standard metric to evaluate language models (e.g., Merity et al. (2017), Lample and Conneau (2019)). However, it is important to mention that this metric has some limitations when comparing language models with different vocabularies (Chen et al., 1998) and it does not necessarily reflect the learned linguistic features (Meister and Cotterell, 2021).

3 Methodology

3.1 Language model

As previously mentioned, in this study, we use Albertina PT-* (Rodrigues et al., 2023) publicly available on the Hugging Face platform [1]. Albertina is a BERT-based large language model with 900M parameters, 24 layers, and a hidden size of 1,536.

Albertina PT-PT was trained over a 2.2 billion token data set which is composed of some openly available corpora of European Portuguese: OSCAR

[1] https://huggingface.co/PORTULAN

[2], DCEP [3], Europarl [4], and ParlamentoPT [5].

Albertina PT-BR was trained over the 2.7 billion token BrWac data set (Wagner Filho et al., 2018).

As both Albertina PT-PT and Albertina PT-BR have the same number of parameters, layers, and hidden sizes, they are adapted for this comparative study.

3.2 Test Data

We consider in this study the following Portuguese varieties spoken in Africa: Angola, Cape Verde, Guinea-Bissau, Mozambique, and São Tomé and Príncipe[6].

The texts in our test sets were extracted from the Corpus Africa which is a subset of the Reference Corpus of Contemporary Portuguese (CRPC). The CRPC is a large electronic corpus of European Portuguese and other varieties. It encompasses 311,4 million words and covers several types of written texts (literary, newspaper, technical, etc.) and spoken texts (formal and informal)[7]. We extracted the sentences without any restriction regarding genre of text.

Variety	Code	Sentences	
		Spoken	Written
Angola	pt-AO	1,776	1,792
Cape Verde	pt-CV	1,794	1,794
Guinea-Bissau	pt-GW	941	1,800
Mozambique	pt-MO	790	1,800
São Tome and Principe	pt-ST	1,277	1,800

Table 1: Languages in the test set.

Table 2 shows the languages in the test set[8]. The number of tokens of each variety-specific corpus is presented in Annex A.

In the extraction process, the aim was to have 1,800 sentences per corpus (randomly selected from the ones in the CRPC). In some cases, the number of available sentences was inferior to this number, and when processing with the language models, some were excluded due to an excessive number of tokens (maximum sequence length of 512 for both models).

3.3 Evaluation

For each test corpus (i.e., written and spoken for each variety of Portuguese), we calculate the negative log-likelihood (NLL) loss for each sentence using both Albertina PT-PT and Albertina PT-BR.

Then, we compute the perplexity of each sentence, which is a measure of how well the language model predicts the given sequence. To do so, we take the mean of the modified negative log-likelihood values and then exponentiate the result.

Finally, we calculate for each corpus the percentage of sentences where the Albertina PT-PT presented a lower value of NLL.

4 Results

Table 2 presents the results obtained for each corpus set in terms of the percentage of sentences in the corpus where Albertina PT-PT performs better than Albertina PT-BR (i.e., where the perplexity measure of Albertina PT-PT is lower than the one obtained with Albertina PT-BR).

Variety	Code	% PT-PT > PT-BR	
		Spoken	Written
Angola	pt-AO	71.1	79.7
Cape Verde	pt-CV	77.4	71.4
Guinea-Bissau	pt-GW	77.8	55.0
Mozambique	pt-MO	72.2	53.7
São Tome and Principe	pt-ST	76.2	60.2

Table 2: Percentage of sentences where perplexity value of Albertina PT-PT is lower than Albertina PT-BR for each test set.

It is possible to notice that the Albertina PT-PT model tends to perform better in comparison with Albertina PT-BR for all languages regarding perplexity measures.

This advantage of the European Portuguese model is more accentuated for the spoken corpora where in more than 70% of the sentences Albertina PT-PT provided lower values of perplexity. On the other hand, concerning the written corpora, only for pt-AO and pt-CV the percentage was higher

[2] https://huggingface.co/datasets/oscar-corpus/OSCAR-2301

[3] https://joint-research-centre.ec.europa.eu/language-technology-resources/dcep-digital-corpus-european-parliament_en

[4] https://www.statmt.org/europarl/

[5] https://huggingface.co/datasets/PORTULAN/parlamento-pt

[6] The variety spoken in Equatorial Guinea was not evaluated due to the lack of available corpus in the CRPC

[7] https://clul.ulisboa.pt/en/projeto/crpc-reference-corpus-contemporary-portuguese

[8] data retrieved from the Reference Corpus of Contemporary Portuguese (CRPC) of the Centre of Linguistics of the University of Lisbon - CLUL (version 3.0 2012, using CQP-Web in the period [10/2023])

than 70. For pt-GW and pt-MO, the difference between the two models is much less pronounced.

For a better analysis of these results, we decided to test both models with Brazilian and Portuguese texts. The idea is to check if the adjustability of the model to a certain variety can be measured with the proposed methodology.

Thus, we extracted 448 sentences (11,611 tokens) from the CRPC for the Brazilian variety of Portuguese and 500 sentences (5,985 tokens) regarding the European Portuguese, then, we proceeded with the same analysis that was conducted for the African varieties. Only written language was considered. The results are presented in Table 3.

Variety	Code	% PT-PT > PT-BR
Brazil	pt-BR	51.1
Portugal	pt-PT	73.6

Table 3: Percentage of sentences where perplexity value of Albertina PT-PT is lower than Albertina PT-BR for European and Brazilian varieties of Portuguese (written register).

The obtained results show that while the Albertina PT-PT seems well adapted to the European variety of Portuguese, regarding the Brazilian texts, the results do not indicate that one model outperforms the other.

5 Discussion

The idea to conduct this general comparative analysis of the performance of language models regarding different varieties of Portuguese is due to the lack of available corpora for each variety that would enable more specific extrinsic examination.

The results presented in Table 2 indicate that the Albertina PT-PT model seems to perform better than the PT-BR model for the African varieties of Portuguese, except for texts in the written register coming from Guinea-Bissau and Mozambique. Regarding Angola and Cape Verde, results were closer to the ones obtained with European Portuguese texts.

Regarding the result obtained for texts in pt-BR (Table 3), both models perform similarly. This can be due to the lack of control concerning genre in this study. The BrWac data-set used to train Albertina PT-BR is a Web corpus, while the test sentences we extracted come from magazines, newspapers, and books. Moreover, the pt-BR test set is composed of texts from 1950 to 2000, a factor that can also have influenced the results. Therefore, before using one model instead of the other just regarding the language variety, one must also check if its training data corresponds to the intended usage.

In this study, we have not analysed the impact of the New Agreement Spelling of the Portuguese Language of 1990[9]. As the selected data-sets may contain texts prior to this agreement, results may have been influenced by this.

Although the results show that the Albertina PT-PT model tends to perform better for the African varieties of Portuguese, this does not mean that this model is well-adapted to be used in downstream NLP tasks for them. Instead, the development of specific models for each variety should be considered for the overall improvement of the NLP results of Portuguese.

Since perplexity measure has some limitations when comparing language models with different vocabularies, we decided to complete our analysis by examining the performance of Albertina PT-PT and PT-BR for part-of-speech tagging. This study is possible as the CRPC also provides POS labels.

Thus, we composed for each African variety of Portuguese and for the European one train and test sets composed of 800 and 200 sentences respectively. We used this data to train and test LSTM models[10] and we added the Albertina embeddings as the first layer. We also tested without the added embeddings to create a baseline.

The results, in terms of accuracy, of the POS-tagging task are presented in tables 4 and 5 for written and spoken texts respectively[11].

It is possible to notice that in almost all cases, the addition of the embeddings in the first layer of the LSTM tends to improve overall accuracy. However, we did not conduct any statistical validation to check whether the improvements are statistically relevant or not.

The POS-tagging results show that, although Albertina PT-PT presented better perplexity measures for African varieties of Portuguese, when this model is applied for this specific NLP task, it does not outperform Albertina PT-BR, even when tested with the European Portuguese corpus.

These unexpected results confirm that further

[9] https://www.priberam.pt/docs/AcOrtog90.pdf

[10] LSTM parameters: epochs=5, batch size=32, validation split=0.2.

[11] For European Portuguese, we only tested with written texts as CRPC does not have spoken ones for this variety

Variety code	No embeddings	Albertina PT-PT	Albertina PT-BR
pt-AO	87.12	87.40	**87.58**
pt-CV	88.15	87.99	**88.88**
pt-GW	84.16	84.55	**86.91**
pt-MO	94.41	96.09	**96.14**
pt-ST	80.89	86.15	**86.78**
pt-PT	91.96	93.39	**93.62**

Table 4: Accuracy of the POS-tagging task for written texts.

Variety code	No embeddings	Albertina PT-PT	Albertina PT-BR
pt-AO	93.69	94.61	**95.20**
pt-CV	92.61	**94.96**	94.92
pt-GW	93.04	94.51	**94.57**
pt-MO	**88.94**	88.90	88.92
pt-ST	94.42	94.66	**95.38**

Table 5: Accuracy of the POS-tagging task for spoken texts.

more specific analysis should be conducted regarding African varieties of Portuguese as the performance of the language models may vary strongly depending on the task.

6 Conclusion and Future Work

In this paper, we presented a comparative evaluation, regarding African varieties of Portuguese, of two state-of-the-art language models, one trained on European Portuguese (Albertina PT-PT), and the other (Albertina PT-BR), on the Brazilian variety of this language.

For each variety, we composed two test sets (spoken and written language) and we calculated the percentage of sentences where the Albertina PT-PT model presented a lower perplexity score when compared to Albertina PT-BR. The idea was to check whether one model is more adapted than the other for the African varieties of Portuguese as, until today, there is no specific language model trained specifically for them.

The obtained results show that Albertina PT-PT seems to outperform Albertina PT-BR in all scenarios regarding the spoken corpora. However, in the written register, the superiority of Albertina PT-PT is less evident for the Portuguese varieties of Guinea-Bissau, Mozambique, and, to a lesser extent, São Tomé and Principe. Moreover, we conducted the same analysis with written texts regarding European and Brazilian Portuguese. As expected, Albertina PT-PT seems more adapted for the European variety, however regarding Brazilian Portuguese, both models performed equally. This can be due to discrepancies between the training data used to create Albertina PT-BR and our test-set.

However, when these models were used in the specific task of part-of-speech tagging, we showed that Albertina PT-BR outperforms PT-PT in almost all cases, even for POS labeling of European Portuguese texts.

The obtained results regarding perplexity and POS-tagging show that there is still a lot of work to be conducted to understand how well existing Portuguese language models perform with African varieties of Portuguese.

Therefore, one perspective for future work is to conduct this analysis in a more controlled scenario regarding the test-sets. Ideally, more sentences should be considered for a complete statistical analysis of the results. Furthermore, as attention to global varieties of Portuguese increases, we hope to see more datasets become available for downstream tasks in these varieties, upon which we can experiment with.

7 Ethics Statement

We affirm our commitment to conducting ethical research and have considered the broader societal implications of our work. We also respect copyright laws and intellectual property rights, giving proper attribution to the works of others in our research.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

António Branco, Sara Grilo, and João Silva. 2022. European language equality - d1.28 - report on the portuguese language.

Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. 1998. Evaluation metrics for language models.

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. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Pablo Botton da Costa, Matheus Camasmie Pavan, Wesley Ramos dos Santos, Samuel Caetano da Silva, and Ivandr'e Paraboni. 2023. BERTabaporu: assessing a genre-specific language model for Portuguese NLP. In *Recents Advances in Natural Language Processing (RANLP-2023)*, Varna, Bulgaria.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas, TX, USA.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

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-*.

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: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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

A Number of tokens in test sets

Variety	**Code**	**Tokens**	
		Spoken	**Written**
Angola	pt-AO	31,892	49,226
Cape Verde	pt-CV	23,515	53,136
Guinea-Bissau	pt-GW	25,963	41,306
Mozambique	pt-MO	31,839	33,007
São Tome and Principe	pt-ST	25,703	22,971