CardioBERTpt: Transformer-based Models for Cardiology Language Representation in Portuguese

1st Elisa Terumi Rubel Schneider
Pontifícia Universidade Católica do Paraná
Curitiba, Brazil
0000-0002-8921-5598

2rd Yohan Bonescki Gumiel
Pontifícia Universidade Católica do Paraná
Heart Institute - InCor/HC FMUSP
Curitiba, Brazil
0000-0001-8239-2930

3nd João Vitor Andrioli de Souza
Comsentimento
Curitiba, Brazil
0000-0002-8950-0890

4rd Lilian Mie Mukai
Pontifícia Universidade Católica do Paraná
Heart Institute - InCor/HC FMUSP
Curitiba, Brazil
0000-0002-6075-3560

5rd Lucas Emanuel Silva e Oliveira
Comsentimento
Curitiba, Brazil
0000-0003-1811-5087

6rd Marina de Sa Rebelo
Heart Institute-InCor/HC FMUSP
São Paulo, Brazil
0000-0001-6069-2529

7rd Marco Antonio Gutierrez
Heart Institute - InCor/HC FMUSP
São Paulo, Brazil
0000-0003-0964-6222

8rd Jose Eduardo Krieger
Heart Institute - InCor/HC FMUSP
São Paulo, Brazil
0000-0001-5464-1792

9rd Douglas Teodoro
University of Geneva
Geneva, Switzerland
0000-0001-6238-4503

10th Claudia Moro
Pontifícia Universidade Católica do Paraná
Curitiba, Brazil
0000-0003-2637-3086

11th Emerson Cabrera Paraiso
Pontifícia Universidade Católica do Paraná
Curitiba, Brazil
0000-0002-6740-7855

***Abstract*—Contextual word embeddings and the Transformers architecture have reached state-of-the-art results in many natural language processing (NLP) tasks and improved the adaptation of models for multiple domains. Despite the improvement in the reuse and construction of models, few resources are still developed for the Portuguese language, especially in the health domain. Furthermore, the clinical models available for the language are not representative enough for all medical specialties. This work explores deep contextual embedding models for the Portuguese language to support clinical NLP tasks. We transferred learned information from electronic health records of a Brazilian tertiary hospital specialized in cardiology diseases and pre-trained multiple clinical BERT-based models. We evaluated the performance of these models in named entity recognition experiments, fine-tuning them in two annotated corpora containing clinical narratives. Our pre-trained models outperformed previous multilingual and Portuguese BERT-based models for cardiology and multi-specialty environments, reaching the state-of-the-art for analyzed corpora, with 5.5% F1 score improvement in TempClinBr (all entities) and 1.7% in SemClinBr (`Disorder` entity) corpora. Hence, we demonstrate that data representativeness and a high volume of training data can improve the results for clinical tasks, aligned with results for other languages.**

***Index Terms*—natural language processing, transformer, clinical texts, language model**

This work was partly supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by Foxconn Brazil and Zerbini Foundation as part of the research project Machine Learning in Cardiovascular Medicine.

I. Introduction

Non-communicable diseases (NCDs) cause over 70% of all deaths worldwide and require continuous monitoring by various clinical specialties. Extracting information from unstructured text fields in electronic health records (EHRs) [1] using natural language processing (NLP) techniques can improve healthcare delivery, but challenges in data collection and heterogeneity limit their use.

Language models like BERT [2] and GPT-3 [3] improve information extraction from clinical texts and facilitate transfer learning. Pre-training self-supervised models and adapting them to downstream NLP tasks is recommended in healthcare where data annotation is complex. However, there are limited pre-trained Transformer-based models specific to Portuguese language in the medical domain, such as BioBERTpt [4] and GPT2-Bio-PT, particularly in cardiology. To address this gap, we proposed CardioBERTpt, pre-trained specifically on large number of cardiology ambulatory notes to improve performance on NLP tasks for this domain.

As CardioBERTpt can be leveraged for several downstream clinical NLP tasks, we have released it publicly available at https://github.com/HAILab-PUCPR/CardioBERTpt.

2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3503-1224-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CBMS58004.2023.00247

2372-9198/23/$31.00 ©2023 IEEE
DOI 10.1109/CBMS58004.2023.00075

II. Related work

NLP models based on the Transformers architecture are the current state-of-the-art in NLP tasks. Transfer learning is effective for adapting pre-trained models to specific contexts or tasks, such as biomedical and clinical domains [5]. Medical texts present challenges such as different formats; high numbers of biomedical entities, abbreviations and acronyms; and clinical concepts with many synonyms [6].

Several language models were developed specifically for biomedical and clinical domains, as observed in [7]. The first models for the biomedical domain emerged by adjusting the weights of a general-domain BERT (used as checkpoint) to the biomedical domain, without any supervision, like BioBERT [8], ClinicalBERT [9]. On the other hand, some models, as PubMedBERT [10] (with texts from the biomedical domain), were trained from scratch without adapting a generalist model. However, pre-training a model from scratch can be computationally expensive and time-consuming, especially for large models with many parameters. Although PubMedBERT achieved higher results than other adapted biomedical models, BiobERT has reached superior results in two of the 14 tasks analyzed in [10], indicatinh that model adaptation may be more suitable in certain scenarios.

III. Methods

As shown in Figure 1, we pre-trained models from six different checkpoints using our clinical data. The checkpoints (intermediate saved versions of a pre-trained language model during the training process) involved the BERT-based models available for Portuguese, both generic domain and specialized in the clinical area. For each pre-trained model, we fine-tuned them to the NER task with two corpora in the clinical domain, TempClinBr [11], and SemClinBr [12].

A. Pre-training of CardioBERTpt Models

CardioBERTpt models were pre-training with a dataset of 157,929 de-identify clinical narratives from cardiology ambulatory, whith more than 61 million tokens (264,601 Unique Tokens) and 7 million sentences. The study obtained ethical approval (n. 5944847) from the Institutional Review Board

The models were trained from checkpoints, which involved utilizing weights from another model to initiate the training process, as adopted in English biomedical reference models BioBERT [8] and ClinicalBERT [9]. We trained from six checkpoints of three main models: i) BERTimbau [13], a model pre-trained with Portuguese journalistic data; ii) BERT multilingual (mBERT) [2] pre-trained on texts from Wikipedia for 104 languages (*cased* and *uncased* checkpoints); and iii) BioBERTpt [4], which was pre-trained from BERT multilingual and fine-tuned with Portuguese biomedical and clinical data (*all*, *clin* and *bio* checkpoints). The pre-training parameters are: 15 Epoch; 1e-5 of Learning Rate; Batch Size of 4; Sequence Length of 512; and Mask Probability of 15%.

B. Fine-tuning for Named Entity Recognition

To evaluate the models, we performed two experiments by fine-tuning the models to the clinical NER task, using the TempClinBr and SemClinBr corpora. There is any overlap between training and evaluation corpora.

TempClinBr is a corpus composed of cardiology specialty texts with 126 ambulatory notes. This corpus is labeled to extract entities, temporal expressions, and their temporal relations. In our evaluation, we considered all clinical entity types available in the benchmark dataset, i.e., *Problems*, *Treatments*, *Tests*, *Occurrences*, *Evidence*, and *Clinical Departments*.

SemClinBr is a semantically annotated corpus from different clinical specialties and kinds of medical and nursing notes, such as evolution, discharge summary, and surgical reports. This corpus is formed by 1,000 clinical notes written in Portuguese, using three semantic types of UMLS - *Disease or Syndrome*, *Finding*, and *Sign or Symptom* - and the semantic group *Disorder* (created by merging together the other three entities).

We trained a model for each class in the SemClinBr corpus as each term can have more than one UMLS class (multi-label classification). We applied the holdout with a corpus split of 60% for training, 20% for validation, and 20% for test, with the parameters: 10 Epoch; 3e-5 of Learning Rate; Batch Size of 8; Sequence Length of 512; Weight Decay of 0.01; Linear Schedule Warmup of 0.1; and AdamW Optimizer.

C. Evaluation

The evaluation metric used was the micro F1-score. In addition to performing the NER experiment with the CardioBERTpt models, we also assess the original BERTimbau, BioBERTpt, and mBERT as baseline models to compare the results and to measure the impact of using clinical narratives to pre-train from the checkpoints. Therefore, the NER evaluation was used as an extrinsic evaluation of the pre-trained models.

IV. Results and Discussion

Table I presents the F1-score for each named entity and BERT model used in our evaluation. The results show that CardioBERTpt as the base model achieves the highest F1-scores for all the assessed entities on both benchmarks. Moreover, in 34 out of 36 (94%) of the cases the CardioBERTpt models improved the F1-score over the baseline models. The CardioBERTpt model achieved a 5.5% F1-score relative improvement for all entities in TemClinBr benchmark compared to the best baseline, and 3.8% improvement for the entity `Problems`. In the SemClinBr corpus, CardioBERTpt achieved a 3.4% improvement for `Disease`, 7.5% for `Finding`, 7.7% for `Sign`, and 1.7% for `Disorder` compared to the best baseline. The pre-trained models of CardioBERTpt using BioBERTpt(all) and multilingual BERT-cased achieved the best F1-score for two named entities each and the pre-trained models of CardioBERTpt with BERTimbau and multilingual BERT-uncased achieved the best F1-score for one named entity each.

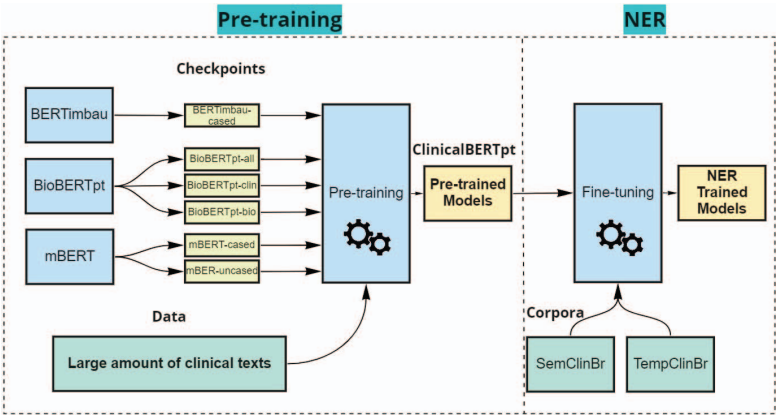

Fig. 1. General view of the proposed method. A large amount of clinical texts were fed to pre-trained checkpoints to create our clinical pre-trained models. These models are then used to extract information from two Portuguese clinical NER corpora, SemClinBr and TempClinBr.

TABLE I
THE AVERAGE F1-SCORES OF THE NER EXPERIMENTS, FOR EACH MODEL EVALUATED. IN BOLD, THE BEST RESULTS.

	TempClinBr		SemClinBr			
Model	**All entities**	**Problem**	**Disease**	**Sign**	**Finding**	**Disorder**
Baselines						
BERTimbau-cased	0.7268	0.7340	0.5014	0.5378	0.5073	0.5906
BioBERTpt(all)	0.8032	0.8204	0.5549	0.5420	0.5051	0.6154
BioBERTpt(bio)	0.7552	0.7904	0.4951	0.5462	0.5158	0.5965
BioBERTpt(clin)	0.7971	0.7889	0.5776	0.5433	0.5095	0.6161
mBERT-cased	0.7374	0.7559	0.5291	0.5213	0.5002	0.6012
mBERT-uncased	0.6646	0.7057	0.4894	0.5328	0.4685	0.5932
Our pre-trained models						
CardioBERTpt-BERTimbau-cased	0.8145	0.8273	0.5761	**0.5880**	0.5366	0.6213
CardioBERTpt-BioBERTpt(all)	0.8332	0.8110	**0.5975**	0.5565	0.5260	**0.6267**
CardioBERTpt-BioBERTpt(bio)	0.8357	0.8188	0.5527	0.5527	0.5510	0.6198
CardioBERTpt-BioBERTpt(clin)	0.8285	0.8467	0.5706	0.5614	0.5466	0.6207
CardioBERTpt-mBERT-cased	**0.8470**	0.8309	0.5632	0.5713	**0.5545**	0.6214
CardioBERTpt-mBERT-uncased	0.8396	**0.8517**	0.5914	0.5637	0.5233	0.6222

A non-parametric Friedman test was conducted to determine the statistical significance of the results. The obtained *p value* was below the significance level (0.05), indicating significant differences among the models. Subsequently, a Nemenyi post-hoc test was performed to identify the pairs of groups that exhibited significant differences. Figure 2 illustrates that both CardioBERTpt-mBERT-cased and CardioBERTpt-mBERT-uncased outperformed the baseline models (mBERT-cased, BERTimbau-cased, and mBERT-uncased) with statistical significance. Additionally, CardioBERTpt-BERTimbau-cased, CardioBERTpt-BioBERTpt(all), and CardioBERTpt-BioBERTpt(clin) models achieved statistically superior results compared to mBERT-uncased.

Considering only the baseline, without further pre-training, the BioBERTpt models achieved the best results, as expected, since they are in-domain models. However, the CardioBERTpt models trained from BioBERTpt (”all”, ”bio” and ”clin” versions) had the best result only for the `Disease` entity from TempClinBr corpus and for the `Disorder` entity from SemClinBr corpus. Moreover, the CardioBERTpt models trained from BioBERTpt(clin) and BioBERTpt(bio) did not perform better in any entity. This was unexpected since the models from BioBERTpt were pre-trained using clinical and biomedical data. We believe that the improvement of out-of-domains models over BioBERTpt is due to the more significant amount of data used for pre-training CardioBERTpt and the cardiology domain, which might represent better the narratives contained in TempClinBr and SemClinBr corpora. The data we used in CardioBERTpt has 2.2 times more tokens compared to the data used on the training of BioBERTpt(clin) model and 1.4 times more tokens than BioBERTpt(all) model.

We also analyzed the output of the NER model for the named entities `Problems` in Figure 3. `Problems` is often one of the most relevant for hospitals, as it enables the organization of clinical notes using a problem-oriented medical record approach. he CardioBERTpt-trained model had high precision considering compound terms, although it missed three entities. Our models extracted detailed clinical information, including location and degree of conditions like ”systolic murmur” and ”left renal stenosis.” Specificity and granularity were prioritized in extracting qualifying values such as ”$<$ 50%”.

CardioBERTpt shows its capability to utilize pre-trained medical language knowledge to enhance the identification

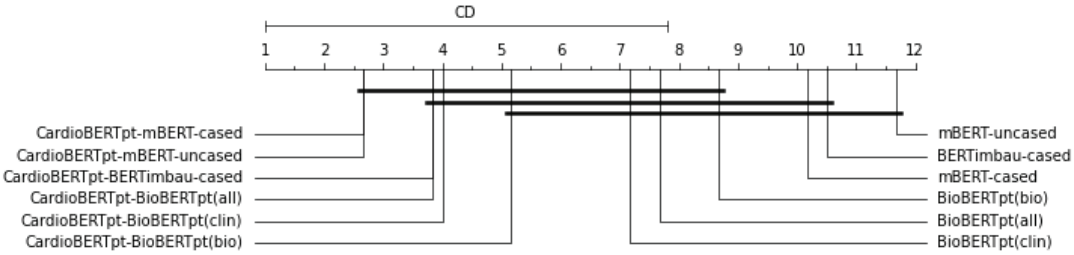

Fig. 2. Results of the Friedman test and Nemenyi post-hoc test with p = 0.05

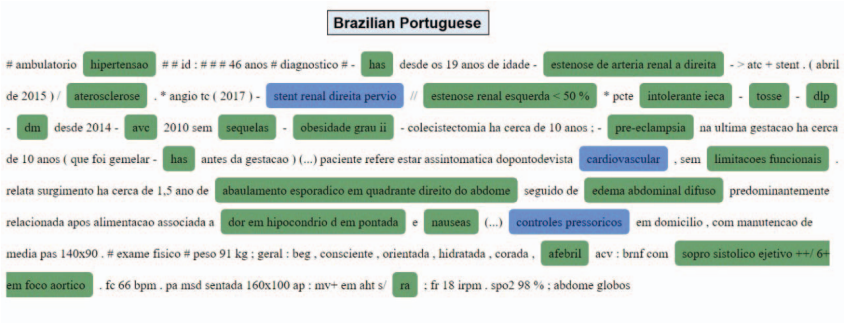

Fig. 3. Prediction examples in a real clinical text from TempClinBr by our clinical model. In green, entities of type Problem correctly detected, and in blue, not annotated entities detected by the model. Translation in: https://github.com/HAILab-PUCPR/CardioBERTpt/tree/main

and classification of medical entities, despite the evaluation datasets are not exclusively from cardiology. The extent of improvement varies for each entity type, potentially due to variations in frequency and complexity within the dataset. These findings highlight the advantages of employing pre-trained models for specific domains. CardioBERTpt is release publicly[1].

REFERENCES

[1] Reading Turchioe, M., Volodarskiy, A., Pathak, J., Wright, D. N., Tcheng, J. E., & Slotwiner, D. (2021). "Systematic review of current natural language processing methods and applications in cardiology", Heart. Advance online publication. https://doi.org/10.1136/heartjnl-2021-319769

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

[3] Tom B. Brown et al. (2020). "Language Models are Few-Shot Learners", In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems, 33, pp. 1877-1901. Curran Associates, Inc.

[4] Elisa Terumi Rubel, et al. "BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition", Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Nov. 2020, pp. 65-72, doi: 10.18653/v1/2020.clinicalnlp-1.7.

[5] Laparra E, Mascio A, Velupillai S, Miller T. "A Review of Recent Work in Transfer Learning and Domain Adaptation for Natural Language Processing of Electronic Health Records.", Yearb Med Inform. 2021 Aug;30(1):239-244. doi: 10.1055/s-0041-1726522. Epub 2021 Sep 3. PMID: 34479396; PMCID: PMC8416218.

[6] Lynda Tamine and Lorraine Goeuriot. "Semantic Information Retrieval on Medical Texts: Research Challenges, Survey, and Open Issues". ACM Comput. Surv. 54, 7, Article 146 (September 2022), 38 pages. https://doi.org/10.1145/3462476

[7] Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2022). "AMMU: A survey of transformer-based biomedical pretrained language models", Journal of Biomedical Informatics, 126, 103982. https://doi.org/10.1016/j.jbi.2021.103982

[8] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., & Kang, J. (2019). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", Bioinformatics, 36(4), 1234-1240. doi: 10.1093/bioinformatics/btz682

[9] Alsentzer, E., Murphy, J., Boag, W., Weng, W., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72-78). Association for Computational Linguistics. doi: 10.18653/v1/W19-1909

[10] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. (2022). "Language Model Pretraining for Biomedical Natural Language Processing", ACM Trans. Comput. Healthcare 3(1), 1–23. doi: 10.1145/3458754

[11] TempClinBr, "https://github.com/HAILab-PUCPR/TempClinBr", 2023.

[12] Lucas Emanuel Silva e Oliveira et al. (2022) "SemClinBr - A Multi-Institutional and Multi-Specialty Semantically Annotated Corpus for Portuguese Clinical NLP Tasks", Journal of Biomedical Semantics 13 (1). doi: 10.1186/s13326-022-00269-1.

[13] Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo. "BERTimbau: Pretrained BERT Models for Brazilian Portuguese." In Intelligent Systems, edited by Ricardo Cerri and Ronaldo C. Prati, 403-417. Springer International Publishing, 2020.

[1]https://github.com/HAILab-PUCPR/CardioBERTpt