# DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains

Yanis Labrak*[1,4] Adrien Bazoge*[2,3] Richard Dufour[2] Mickael Rouvier[1]
Emmanuel Morin[2] Béatrice Daille[2] Pierre-Antoine Gourraud[3]

(1) LIA, Avignon Université (2) LS2N, Nantes Université (3) Clinique des données, CHU de Nantes (4) Zenidoc

# Summary

I. Language Modeling in Healthcare

II. Comparison of pre-training strategies, data sources and sizes

III. Evaluation of 13 models on 11 tasks

IV. Distribution of NACHOS and DrBERT

# Language Modeling

- Transformer-based approaches, such as BERT, offer **huge performance gain** on a lot of NLP tasks
- Has been adapted to **French** with **CamemBERT** and **FlauBERT**
- **On medical tasks, domain-specific models** in English raised the bar even higher
    - PudMedBERT, BioBERT, ClinicalBERT and other
- Languages others than English are rarer and rely primarily on **continual pre-training** using an existing generic model
- Unlike generic models, **no open-source** model is available for **biomedical domain in French** yet
- BERT-based domain specific model for French **should increase performance** on medical tasks

# Comparison of pre-training strategies and data sources

- **Evaluation of the impact of public and private medical data sources on comparable data sizes**
  - NACHOS: A **1.1B words open-source dataset** of heterogeneous data crawled from **diverse medical domains**, **natures** and **styles**
  - NBDW: A private dataset of sentences taken from 1.7M **anonymized medical records** extracted from the Nantes University Hospital data warehouse

| Corpus | Size | #words | #sentences |
|---|---|---|---|
| $\text{NACHOS}_{large}$ (pub.) | 7.4 GB | 1.1 B | 54.2 M |
| $\text{NACHOS}_{small}$ (pub.) | 4 GB | 646 M | 25.3 M |
| $\text{NBDW}_{small}$ (private) | 4 GB | 655 M | 43.1 M |
| $\text{NBDW}_{mixed}$ (both) | 4+4 GB | 1.3 B | 68.4 M |

- **Comparison of learning strategies**
  - *From scratch* with full model construction
  - *Continual pre-training* using an existing pre-trained model (here, CamemBERT, a French generic model, and PubMedBERT, an English-based medical one)

| Model name | Strategy | Corpus |
|---|---|---|
| DrBERT | From scratch | $\text{NACHOS}_{large}$ |
| DrBERT | From scratch | $\text{NACHOS}_{small}$ |
| ChuBERT | From scratch | $\text{NBDW}_{small}$ |
| ChuBERT | From scratch | $\text{NBDW}_{mixed}$ |
| CamemBERT | continual pre-training | $\text{NACHOS}_{small}$ |
| PubMedBERT | continual pre-training | $\text{NACHOS}_{small}$ |
| CamemBERT | continual pre-training | $\text{NBDW}_{small}$ |

# Evaluation : Data sources and size

- **Performance** evaluation of **13 models** on **11 tasks**, both **public** and **private**
- Our fine-tuned models get **state-of-the-art results** on almost all tasks

| | | aHF NER | aHF CLS | Medical Report NER | Specialities CLS | MUSCA-DET NER | MUSCA-DET CLS | ESSAI POS | CAS POS | FrenchMedMCQA | | QUAERO-EMEA NER | QUAERO-MEDLINE NER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *F1* | *F1* | *F1* | *F1* | *F1* | *F1* | *F1* | *F1* | *Hamming* | *EMR* | *F1* | *F1* |
| General | **CamemBERT OSCAR 138 GB** | 35.13 | 80.13 | 89.35 | 99.20 | 88.54 | 88.20 | 81.10 | 95.22 | 36.24 | **16.55** | 90.71 | 77.41 |
| | **CamemBERT OSCAR 4 GB** | 42.66 | 81.41 | 88.78 | 99.61 | 85.43 | 91.27 | 83.69 | 96.42 | 35.75 | 15.37 | 90.83 | 78.76 |
| | **CamemBERT CCNET 4 GB** | 43.11 | 79.98 | 89.34 | 99.55 | 90.33 | 91.38 | 85.42 | **97.33** | 34.71 | 14.41 | 90.33 | 77.61 |
| Biomedical | **PubMedBERT** | 47.22 | 76.86 | 89.20 | 99.37 | 91.99 | 81.97 | 87.78 | 95.90 | 33.98 | 14.14 | 86.79 | 77.09 |
| | **BioBERT v1.1** | 46.01 | 79.00 | 89.38 | 98.80 | 90.46 | 81.91 | 85.18 | 97.12 | 36.19 | 15.43 | 84.29 | 72.68 |
| | **DrBERT NACHOS**$_{large}$ | 48.22 | 81.25 | **89.83** | **99.86** | 91.04 | **92.24** | **89.75** | 95.65 | 36.66 | 15.32 | **92.09** | 77.88 |
| | **DrBERT NACHOS**$_{small}$ | 45.93 | 79.87 | 89.44 | 99.85 | 91.77 | 88.57 | 88.76 | 95.70 | **37.37** | 13.34 | 91.66 | 78.18 |
| Clinical | **ClinicalBERT** | 44.70 | 77.12 | 88.77 | 98.58 | 90.36 | 82.95 | 88.24 | 96.73 | 32.78 | 14.19 | 84.79 | 75.05 |
| | **ChuBERT NBDW**$_{small}$ | 49.01 | 81.56 | 89.58 | 99.83 | 92.23 | 92.17 | 87.71 | 95.61 | 35.16 | 14.79 | 88.15 | 74.94 |
| | **ChuBERT NBDW**$_{mixed}$ | **49.14** | **81.98** | 89.30 | 99.81 | **92.73** | 91.71 | 85.73 | 96.35 | 34.58 | 12.21 | 90.52 | **78.63** |

# Evaluation : Pre-training strategies

- From scratch vs. continual pre-training on 4GB of data
- Question-answering tasks require **more domain specific knowledge** to be able to work well
- A study of **model stability** shows a **higher inter-run variability** for the CamemBERT-based models **trained using continual pretraining**

| | aHF NER | aHF CLS | Medical Report NER | Specialities CLS | MUSCA-DET NER | MUSCA-DET CLS | ESSAI POS | CAS POS | FrenchMedMCQA | | QUAERO-EMEA NER | QUAERO-MEDLINE NER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *F1* | *F1* | *F1* | *F1* | *F1* | *F1* | *F1* | *F1* | *Hamming* | *EMR* | *F1* | *F1* |
| **CamemBERT OSCAR 138 GB** | 35.13 | 80.13 | 89.35 | 99.20 | 88.54 | 88.20 | 81.10 | 95.22 | 36.24 | **16.55** | 90.71 | 77.41 |
| **CamemBERT OSCAR 4 GB** | 42.66 | 81.41 | 88.78 | 99.61 | 85.43 | 91.27 | 83.69 | 96.42 | 35.75 | 15.37 | 90.83 | 78.76 |
| **CamemBERT CCNET 4 GB** | 43.11 | 79.98 | 89.34 | 99.55 | 90.33 | 91.38 | 85.42 | **97.33** | 34.71 | 14.41 | 90.33 | 77.61 |
| **PubMedBERT** | 47.22 | 76.86 | 89.20 | 99.37 | 91.99 | 81.97 | 87.78 | 95.90 | 33.98 | 14.14 | 86.79 | 77.09 |
| **ClinicalBERT** | 44.70 | 77.12 | 88.77 | 98.58 | 90.36 | 82.95 | 88.24 | 96.73 | 32.78 | 14.19 | 84.79 | 75.05 |
| **BioBERT v1.1** | 46.01 | 79.00 | 89.38 | 98.80 | 90.46 | 81.91 | 85.18 | 97.12 | 36.19 | 15.43 | 84.29 | 72.68 |
| **DrBERT NACHOS$_{large}$** | 48.22 | 81.25 | **89.83** | **99.86** | 91.04 | 92.24 | **89.75** | 95.65 | 36.66 | 15.32 | **92.09** | 77.88 |
| **DrBERT NACHOS$_{small}$** | 45.93 | 79.87 | 89.44 | 99.85 | 91.77 | 88.57 | 88.76 | 95.70 | **37.37** | 13.34 | 91.66 | 78.18 |
| **ChuBERT NBDW$_{small}$** | 49.01 | 81.56 | 89.58 | 99.83 | 92.23 | 92.17 | 87.71 | 95.61 | 35.16 | 14.79 | 88.15 | 74.94 |
| **ChuBERT NBDW$_{mixed}$** | **49.14** | **81.98** | 89.30 | 99.81 | **92.73** | 91.71 | 85.73 | 96.35 | 34.58 | 12.21 | 90.52 | 78.63 |
| **CamemBERT NACHOS$_{small}$** | 16.08 | 69.80 | 66.74 | 99.54 | 80.96 | 78.70 | 80.04 | 92.46 | 32.87 | 13.76 | 71.10 | 57.43 |
| **PubMedBERT NACHOS$_{small}$** | 48.72 | 81.40 | 89.36 | 99.55 | 91.53 | **93.62** | 83.85 | 96.81 | 35.88 | 15.21 | 91.03 | **81.73** |
| **CamemBERT NBDW$_{small}$** | 19.12 | 76.02 | 69.64 | 99.58 | 81.57 | 77.12 | 79.25 | 93.18 | 27.73 | 11.89 | 61.75 | 53.0 |

# Core message

- DrBERT achieves **state-of-the-art** results in **9 downstream French medical-oriented tasks**
  - Surpasses CamemBERT generic model and English-based domain-specific models
  - Confirms utility of training a medical-specific model in French
- **Data sources matters:** training on **heterogeneous** data is important
  - NACHOS is more robust than using private clinical data only
- More data is better, but does not scale well
- Continual pretraining is a more **effective strategy** when based on domain-specific English models
- The DrBERT models, the NACHOS dataset and the **training scripts** are freely available under the **MIT** license

drbert.univ-avignon.fr

# Thank You

**Looking forward to exchange at poster session in Toronto!**

More information on:
drbert.univ-avignon.fr