Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method

Stella Verkijk & Piek Vossen

In a nutshell

- We explain how we anonymized a medical Language Model pre-trained on private information
- This use-case shows that models like this can be safely published when thoroughly anonymized in this way.

What will be discussed

- Language Models (LMs)
- Privacy Risks when publishing a LM
- MedRoBERTa.nl
- Anonymizing MedRoBERTa.nl
- Testing Anonymity
- Results

Language Models (LM)

- State-of-the-art: Successful as a basis for many NLP systems
- MASK learning objective
- Predicts similarity of words by analyzing their contexts
- Use of vocabulary
- Generative function: fill-mask

Example: *‘MASK was diagnosed with Covid-19’*

Privacy risks when publishing a LM

Machine Learning systems can be queried to discover what data was present in their training data (Mireshghallah et al. (2020))

→ Membership Inference

- Property Inference
- Model Inversion

Example: *‘MASK was diagnosed with Covid-19’*

Patient	0.78
Sir	0.65
Biden	0.51

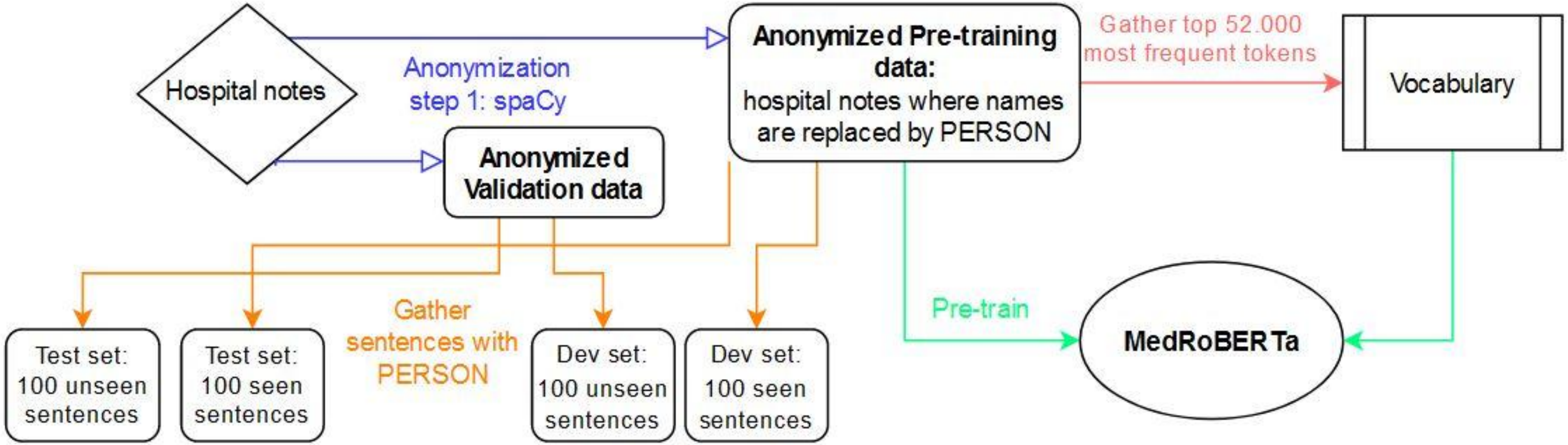
MedRoBERTa.nl (Verkijk & Vossen, 2022)

- pre-trained on secure server
- nearly 10 million hospital notes from Electronic Health Records
- specialized vocabulary

Anonymizing MedRoBERTa.nl

- automatic anonymization of pre-training data
- semi-automatic anonymization of vocabulary

Anonymization phase I: Before pre-training



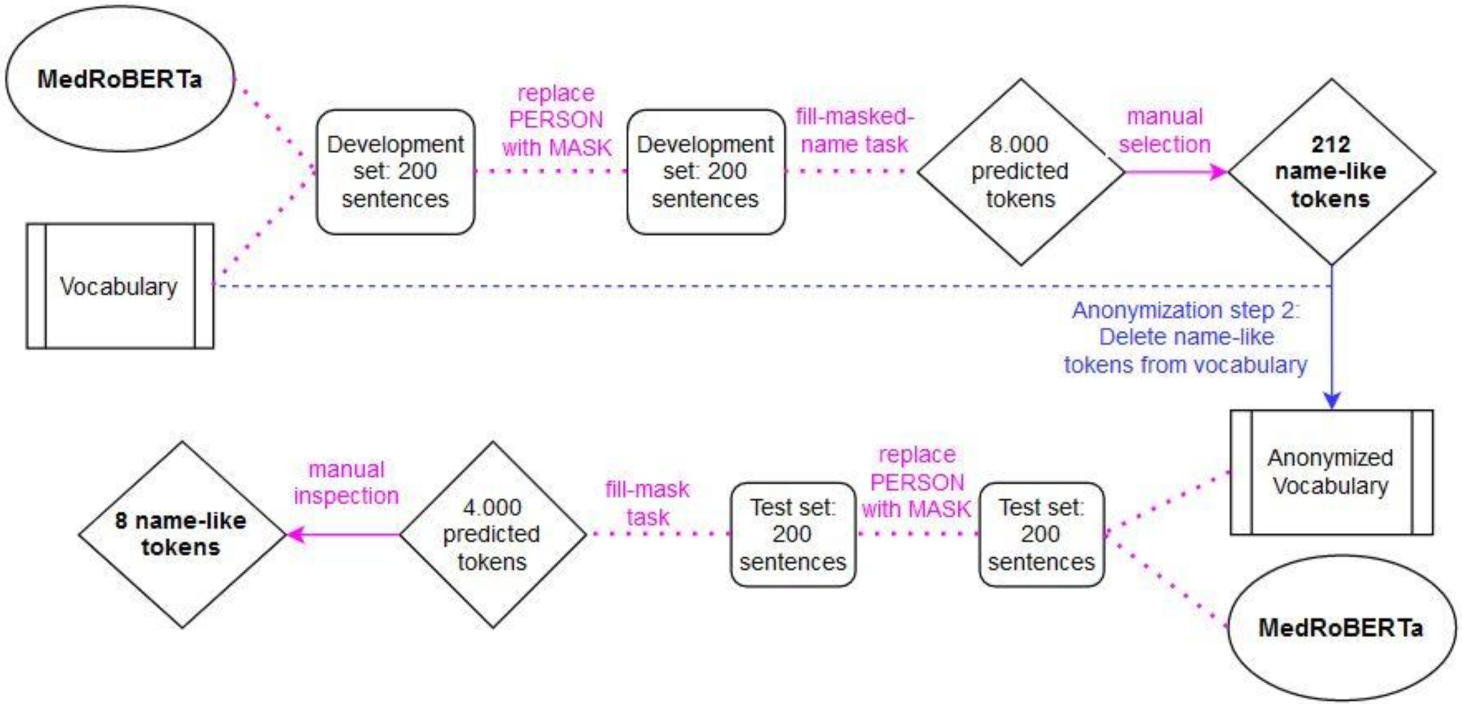
Hospital notes
Anonymization step 1: spaCy
Anonymized Pre-training data:
hospital notes where names are replaced by PERSON
Gather top 52.000 most frequent tokens
Vocabulary
Anonymized Validation data
Gather sentences with PERSON
Pre-train
MedRoBERTa
Test set: 100 unseen sentences
Test set: 100 seen sentences
Dev set: 100 unseen sentences
Dev set: 100 seen sentences

- LM will never learn the names that are replaced by PERSON
- LM will register PERSON as one entity
  → all associations with other names will be pushed to the background

Next step: anonymizing the vocabulary

→ LM will never be able to predict a name that is not in its vocabulary even though it was present in its pre-training data

Anonymization phase II: After pre-training



Testing Anonymity

- Sampled 100 seen & 100 unseen sentences like

  ‘PERSON was diagnosed with Covid-19’ → ‘MASK was diagnosed with Covid-19’

- Performed *fill-masked-name* task and gathered top 20 predictions for each sentence
- Manually inspected the 4.000 predictions for name-like tokens

Results

- 8 name-like tokens in 4.000 predictions
- For 192 out of 200 sentences, no name was ever predicted
- Never the first, most probably prediction (always 6th or later)
- Never the name originally present in the data

MedRoBERTa was published with open access with permission of the Privacy Office of the Amsterdam University Medical Centres.

Resources

MedRoBERTa.nl

https://huggingface.co/CLTL/MedRoBERTa.nl

Code for anonymization
https://github.com/cltl-students/verkijk_stella_rma_thesis_dutch_medical_language_model/tree/master/src/anonymization

Test set of 100 unseen sentences
https://github.com/cltl-students/verkijk_stella_rma_thesis_dutch_medical_language_model/blob/master/src/anonymization/anon_specific_testset_eval_public.csv

Contact

The Computational Linguistics & Text Mining Lab (CLTL)

Vrije Universiteit Amsterdam

(http://www.cltl.nl/)

Stella Verkijk

Researcher at CLTL

stellaverkijk@outlook.com

Piek Vossen

Head of CLTL & Full Professor Computational Lexicology at VU

p.t.j.m.vossen@vu.nl