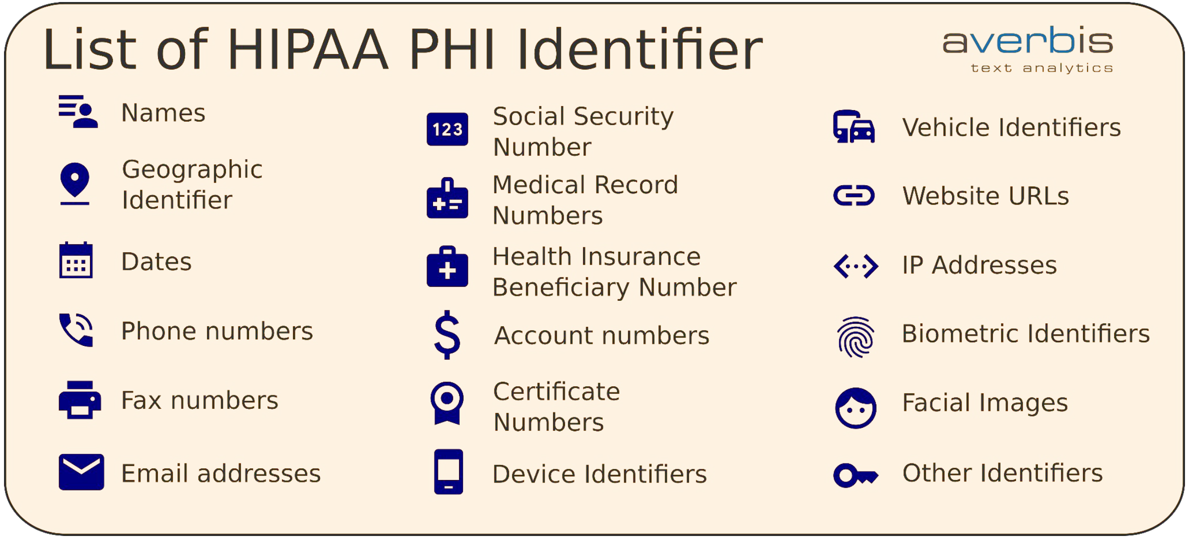
averbis
text analytics

Anonymisation and De-identification of Medical Documents

Averbis GmbH, David Hübner

De-identification: Scope

- **De-identification**: Removal of protected health information (**PHI**)
- In the **EU**, we currently lack an actionable definition of PHI, so we resort to the Health Insurance Portability and Accountability Act (**HIPAA**)



De-identification: Scenario

De-identification: Technical approach

- Use natural language processing (**NLP**) to identify PHI in clinical documents
- NLP models are trained on manually labeled datasets (>175.000 annotations)
- They can identify PHI information based on
  - **Context** (*“Dear Mr …”*)
  - **World knowledge** (*“Berlin”*)
  - **Layout** (to some degree)
- Key technology: Usage of **Transformer**-based architectures, but fast enough to run on-premise (no LLMs)

Record date: 2157-07-17

DEJESUS, JAROME
28044071
07/17/2157

Yoseph Berg, M.D.
Broome Developmental Center
84 Parkhurst Drive
Peculiar, NE 54583

Dear Dr. Berg:

We had the pleasure of caring for your patient, Jarome Dejesus, during his admission to the Erie County Infirmary from 07/16/2157 to 07/17/2157 for evaluation and treatment of his anginal symptoms. As you know, Mr. Dejesus is a very pleasant 74-year-old gentleman with paroxysmal atrial fibrillation who is on Coumadin. He has a history of insulin-dependent diabetes mellitus, hypertension and dyslipidemia. He has known coronary artery disease and underwent stenting of his left anterior descending artery in 05/2155, at which time he had a Taxus drug-eluting stent placed. He returns now with increasing dyspnea on exertion and had an elective cardiac catheterization performed at the Southwest Colorado Medical on 07/13/2157 that showed focal 90% instent restenosis in the left anterior descending artery with no other focally occlusive disease.
He was transferred to the Erie County Infirmary and on the afternoon of 07/16/2157, Dr. Yoseph Flint was able to dilate and stent the lesion in the left anterior descending artery with placement of a 3.5 x 28 mm Cypher drug-eluting stent. The procedure was complicated by brief transient no re-flow that resolved immediately to intracoronary nitroglycerin. The left femoral arteriotomy was closed using an Angio-Seal device.
Mr. Dejesus has spent an uneventful evening on our Cardiovascular Service. He has had no complaints of chest pain or shortness of breath. He is ambulating, eating, and voiding without difficulty.

With warmest regards.

Sincerely,

Irving R. Ramirez, M.D.
ATTENDING

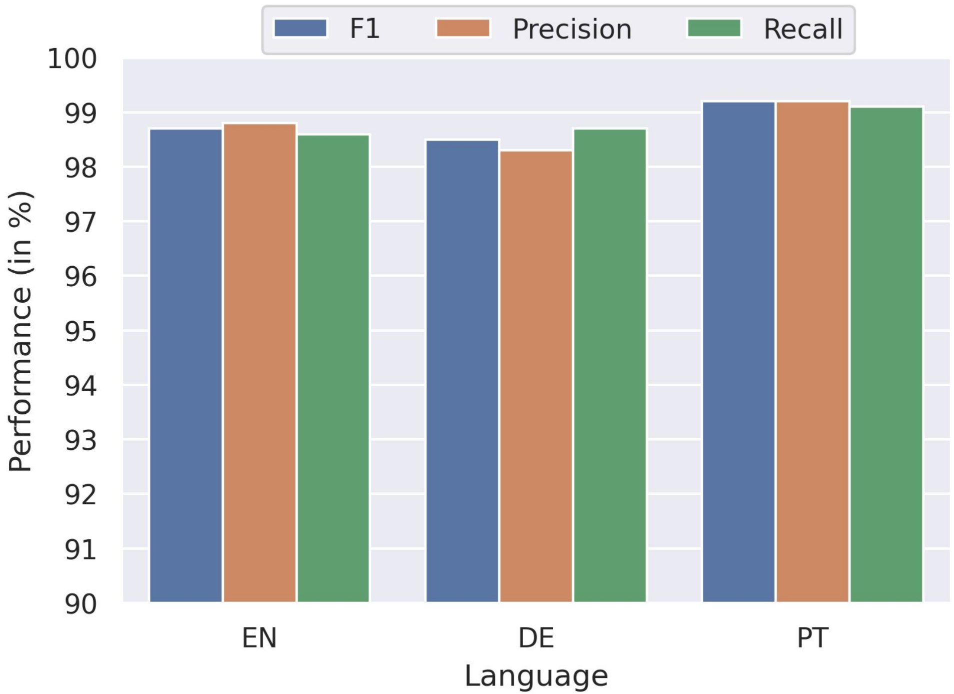
Dictated by: Henry Phelps, PA

eScription document:8-0687422 DFFocus

CC: Robert-Joseph Paige M.D.
Koenen Collins County Clinic
886 Paxton Terrace
Peculiar, NE 19649

De-identification: Key Results

- **Recall > 98%:** We can find 98-99% of all PHI
- **Precision > 98%:** If we mark a Token as PHI, it is correct to an accuracy of over 98%
- A dedicated **Recall-optimized** pipeline achieves an average of **99.3% Recall**



Token-level micro-average evaluation

De-identification: From annotations to de-identified documents

Once PHI was identified, we can modify the original documents.

Common strategies:

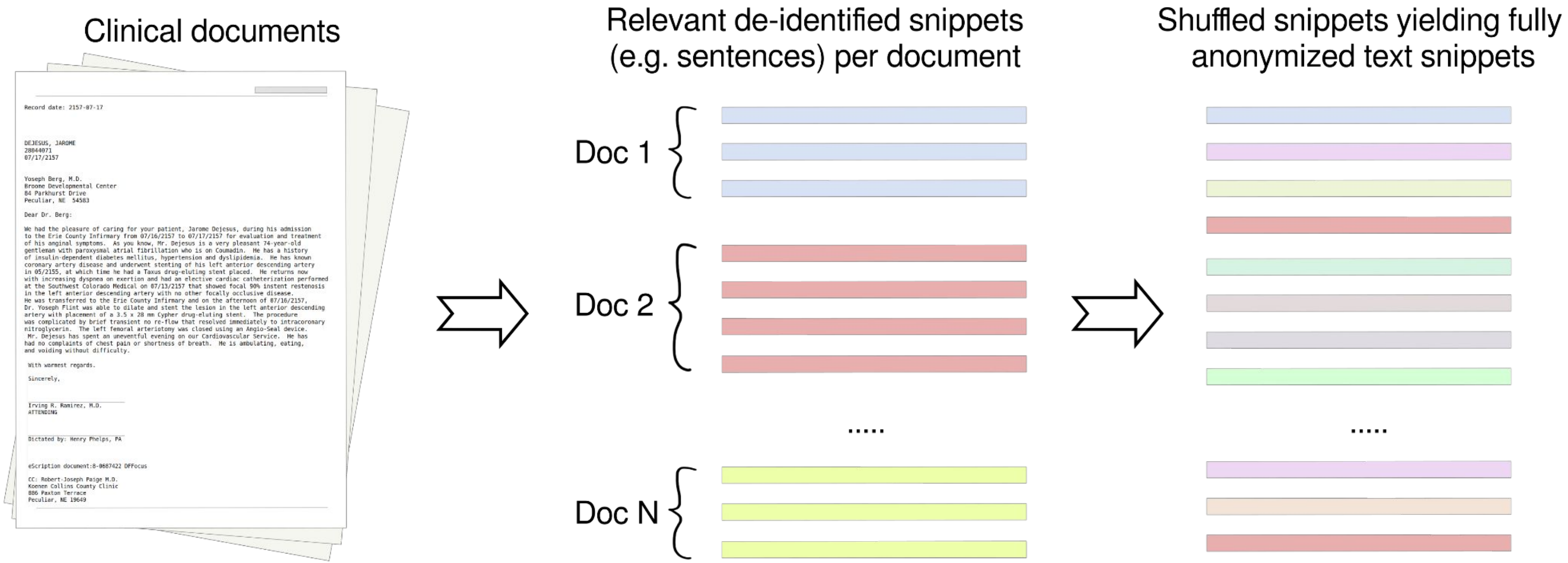
- **Redaction**: Replace PHI by **XXX** or by a tag, e.g. **<date>**
- **Substitution**: Substitute PHI with **synthetic information**
  - *Advantages*: Enhanced security - It is unclear whether any remaining information was originally PHI
  - *Challenge*: Consistent replacements are difficult

De-identification = Anonymization? Not quite.

- Redaction/Substitution of PHI makes it much more difficult to identify the patients
- However, distinct medical diagnoses and contextual cues still may allow to identify the patient (see **k-anonymity**)
- To overcome this, we developed the **Snippet Workflow**

Snippet Workflow for retrieving anonymized texts

Goal: Workflow to retrieve texts that are **fully anonymized** but still valuable for ML model training



averbis
text analytics

David Hübner
Lead ML Engineer at Averbis GmbH
Email: david.huebner@averbis.com