Deidentification

Muhammet ŞANTAŞ

Senior Data Scientist, John Snow Labs

Agenda

1. Introduction
2. Case Study Samples
3. Benchmarks
4. Resources
5. Coding

John Snow LABS

Why we need deidentification?

- Patient privacy
- Data Security
- Research and Analysis
- Compliance with Regulations

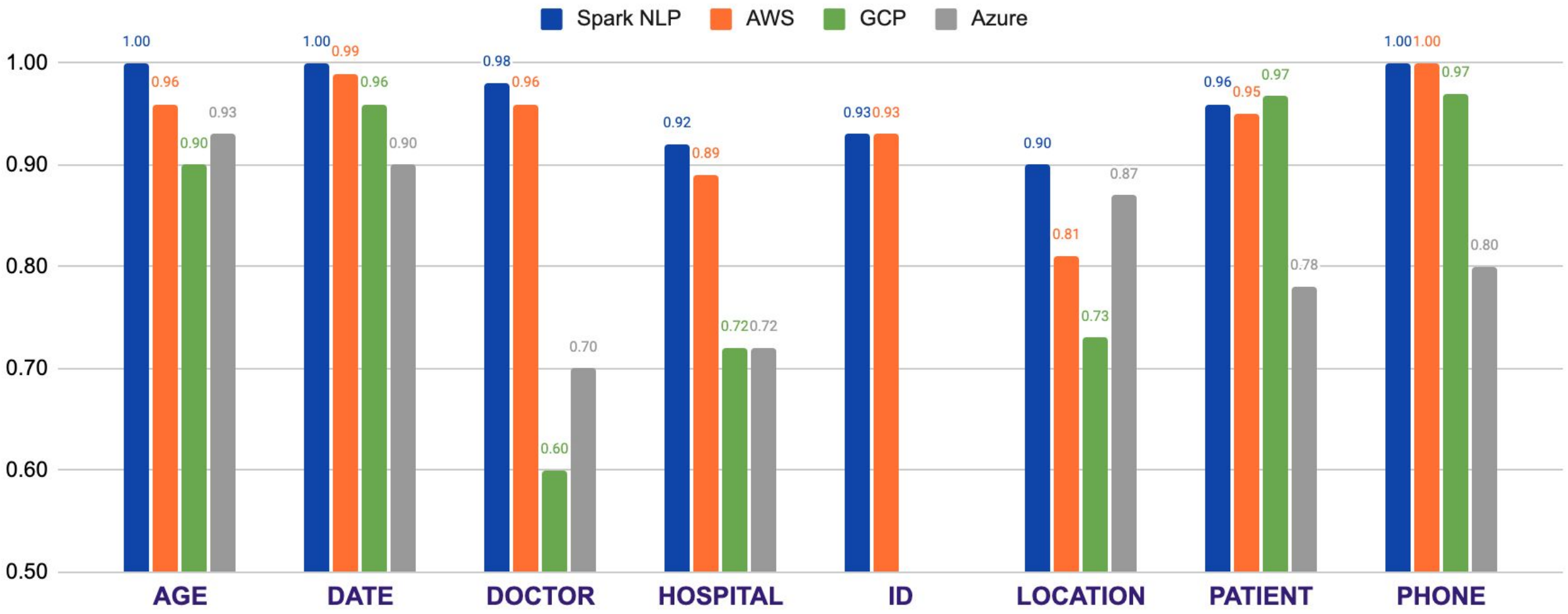
Deidentification Case Studies from the NLP Summit

Lessons Learned De-Identifying 700 Million Patient Notes with Spark NLP

De-Identification Benchmarks (en)

	English	German	French	Spanish	Italian	Portuguese
PATIENT	0.90	0.97	0.94	0.92	0.91	0.95
DOCTOR	0.94	0.98	0.99	0.92	0.92	0.93
HOSPITAL	0.91	1.00	0.94	0.86	0.90	0.90
DATE	0.98	1.00	0.98	0.99	0.98	0.98
AGE	0.94	0.99	0.86	0.98	0.98	0.98
PROFESSION	0.84	1.00	0.81	0.91	0.89	0.90
ORGANIZATION	0.77	0.94	0.77	0.83	0.74	0.97
STREET	0.98	0.98	0.90	0.94	0.98	1.00
CITY	0.83	0.99	0.86	0.84	0.97	0.98
COUNTRY	0.81	0.98	0.90	0.87	0.93	0.91
PHONE	0.94	0.88	0.98	0.90	0.98	0.99
USERNAME	0.92	1.00	0.92	0.74	0.91	0.88
ZIP	0.99	-	1.00	0.99	0.99	0.99

De-Identification Benchmarks (en)



Clinical_Deidentification_SparkNLP_vs_Cloud_Providers_Comparison.ipynb

Deidentification Problem Solving

Entity Extraction (NER)

Create a robust NER pipeline and extract PHI entities.

Merge Entities

Merge the entities coming from NER models by prioritizing.

Deidentification

Adjust the parameters of Deidentification annotator according to the needs.

Out-of-the-Box Deidentification NLP Models

Pretrained Pipelines

index	model	lang
1	clinical_deidentification	de
2	clinical_deidentification	en
3	clinical_deidentification_glove	en
4	clinical_deidentification_glove_augmented	en
5	clinical_deidentification	es
6	clinical_deidentification_augmented	es
7	clinical_deidentification	fr
8	clinical_deidentification	it
9	clinical_deidentification	pt
10	clinical_deidentification	ro
11	clinical_deidentification	ar
12	clinical_deidentification_obfuscation_medium	en
13	clinical_deidentification_obfuscation_small	en
14	clinical_deidentification_generic	en
15	clinical_deidentification_generic_optimized	en
16	clinical_deidentification_langtest	en
17	clinical_deidentification_wip	en
18	clinical_deidentification_subentity_optimized	en
19	clinical_deidentification_subentity_nameAugmented	en
20	clinical_deidentification_subentity	en
21	clinical_deidentification_slim	en
22	clinical_deidentification_obfuscation_small	en
23	clinical_deidentification_obfuscation_medium	en
24	clinical_deidentification_multi_mode_output	en

NER Models (EN)

index	model	lang	index	model	lang	index	model	lang
1	deidentify_dl	en	7	ner_deid_enriched_biobert	en	13	ner_deid_subentity_augmented	en
2	deidentify_large	en	8	ner_deid_generic_augmented	en	14	ner_deid_subentity_glove	en
3	deidentify_rb	en	9	ner_deid_generic_glove	en	15	ner_deid_synthetic	en
4	ner_deid_augmented	en	10	ner_deid_large	en	16	ner_deidentify_dl	en
5	ner_deid_biobert	en	11	ner_deid_sd	en			
6	ner_deid_enriched	en	12	ner_deid_sd_large	en			

NER Models (Other Lang)

index	model	lang	index	model	lang
1	ner_deid_generic	de	14	ner_deid_generic	fr
2	ner_deid_subentity	de	15	ner_deid_subentity	fr
3	ner_deid_generic	es	16	ner_deid_generic	it
4	ner_deid_generic_augmented	es	17	ner_deid_subentity	it
5	ner_deid_generic_roberta	es	18	ner_deid_generic	pt
6	ner_deid_generic_roberta_augmented	es	19	ner_deid_subentity	pt
7	ner_deid_subentity	es	20	ner_deid_subentity	ro
8	ner_deid_subentity_augmented	es	21	ner_deid_subentity_bert	ro
9	ner_deid_subentity_roberta	es	22	ner_deid_generic	ro
10	ner_deid_subentity_roberta_augmented	es	23	ner_deid_generic_bert	ro
11	ner_deid_subentity	ar	24	ner_deid_generic	ar
12	ner_deid_subentity_arabert	ar	25	ner_deid_generic_arabert	ar
13	ner_deid_subentity_camelbert	ar	26	ner_deid_generic_camelbert	ar

De-Identification

	Sentence	Masked	Masked with Chars	Masked with Fixed Chars	Obfuscated
0	Name : Hendrickson, Ora, Record date: 2093-01-13, Age: 25, # 719435.	Name : <PATIENT>, Record date: <DATE>, Age: <AGE>, # <ZIP>.	Name : [**********], Record date: [********], Age: **, # [****].	Name : ****, Record date: ****, Age: ****, # ****.	Name : Kathryn Spillers, Record date: 2093-02-17, Age: 86, # 3%05d.
1	Dr. John Green, ID: 1231511863, IP 203.120.223.13.	Dr. <DOCTOR>, ID<IDNUM>, IP <IPADDR>.	Dr. [********], ID[**********], IP [**********].	Dr. ****, ID****, IP ****.	Dr. Dr Orlena Caprice, IDHS:6828076, IP 444.444.444.444.
2	He is a 60-year-old male was admitted to the Day Hospital for cystectomy on 01/13/93.	He is a <AGE> male was admitted to the <HOSPITAL> for cystectomy on <DATE>.	He is a [*********] male was admitted to the [**********] for cystectomy on [******].	He is a **** male was admitted to the **** for cystectomy on ****.	He is a 2 male was admitted to the BROOKE ARMY MEDICAL CENTER for cystectomy on 10-25-1974.
3	Patient's VIN : 1HGBH41JXMN109286, SSN #333-44-6666, Driver's license no:A334455B.	Patient's VIN : <VIN>, SSN <SSN>, Driver's license <DLN>.	Patient's VIN : [**********], SSN [**********], Driver's license [*********].	Patient's VIN : ****, SSN ****, Driver's license ****.	Patient's VIN : 1AAAA00AAAA111000, SSN 999-37-3986, Driver's license S99934770.
4	Phone (302) 786-5227, 0295 Keats Street, San Francisco, E-MAIL: smith@gmail.com	Phone <PHONE>, <STREET>, <CITY>, E-MAIL: <EMAIL>	Phone [**********], [**********], [**********], E-MAIL: [**********]	Phone ****, ****, ****, E-MAIL: ****	Phone 773 553 069, Bouciña 65, Kankakee, E-MAIL: Emeline@google.com

Out-of-the-Box Deidentification Models

De-Identification - Live Demos & Notebooks

Detect PHI Entities from Deidentification

This demo shows how to deidentify protected health information.

Live Demo | Colab

Deidentify Clinical Notes in Different Languages

This demo shows how to deidentify protected health information in English, Spanish, French, Italian, Portuguese, Romanian, and German (...)

Live Demo | Colab

Consistency on Deidentification

Our De-Identification process shown in this demo ensures data clarity, usability and consistency while prioritizing privacy and (...)

Live Demo | Colab

How to Perform Day Shifting and Normalization for Testing Data

This demo demonstrates to you through the straightforward process of normalizing and shifting dates with ease.

Live Demo | Colab

Detect PHI Entities from Deidentification

Automatically identify demographic information such as **Date, Doctor, Hospital, ID number, Medical record, Patient, Age, Profession**, (...)

Live Demo | Colab

Deidentify structured data

Deidentify PHI information from structured datasets using out of the box Spark NLP functionality that enforces GDPR and HIPP/ (...)

Live Demo | Colab

Deidentify DICOM documents

Deidentify DICOM documents by masking PHI information on the image and by either masking or obfuscating PHI from the metadata. (...)

Live Demo | Colab

De-identify PDF documents - HIPAA Compliance

De-identify PDF documents using HIPAA guidelines by masking PHI information using out of the box Spark NLP models.

Live Demo | Colab

De-identify PDF documents - GDPR Compliance

De-identify PDF documents using GDPR guidelines by anonymizing PHI information using out of the box Spark NLP models. (...)

Live Demo | Colab

Detect PHI Entities from Deidentification (Arabic)

Detect protected health information in Arabic clinical documents using Spark NLP models, identifying up to 17 entities.

Live Demo | Colab

Detect PHI for Generic Deidentification (multilingual)

Deidentification NER is a Named Entity Recognition model that annotates English, German, French, Italian, Spanish, Portugues (...)

Live Demo | Colab

Deidentification Tool

DEMO TOOL
DE-IDENTIFICATION

TEST DATA

Test your data by applying the following available De-identification tools.

Free Text >>
De-identify free text documents

Table >>
Database, XML, JSON, XLSX, CSV

Documents >>
PDF / DOCX / PPTX / JPEG

DICOM >>
De-identify DICOM documents

SELECT LANGUAGE

Auto

Powered by John Snow LABS

Sample Data

Enter Text Below Never Enter Real PHI Data

Harbor Hospital

---

36 Park Avenue, 95108, San Diego, CA, USA
Email: medunites@firsthospital.com,
Phone: (818) 342-7353.

TSICU MRN# 1482928 on 24/06/2019 by ambulance VIN: 1HGBH41JXMN109186.

John Davies is a 62 y.o. patient admitted to ICU after an MVA on 22 Hoyt Street, at 23:00 hours. He works as a driver, and long hours of work reported. He reports dizziness, drowsiness, head ache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. John Davies was seen at 23:12 minutes by attending physician Dr. Meyer Lorand and was scheduled for emergency head and neck CT with further neurological assessment. At 23:18 he was neurologically assessed by Dr. Frank M and was HD stable with normal vital signs and therefore and transferred (ID num 184378) for further radiological investigations.

Other medications:

Test Data

Result Window

Detected language: en

Masked Text | Obfuscated Text

<HOSPITAL>

<STREET>, <ZIP>, <CITY>, <STATE>, <COUNTRY>
Email: <EMAIL>,
Phone: <PHONE>.

TSICU MRN# <MEDICALRECORD> on <DATE> by ambulance VIN:
<VIN>.

<PATIENT> is a <AGE> y.o. patient admitted to ICU after an MVA on <STREET>, at 23:00 hours. He works as a <PROFESSION>, and long hours of work reported. He reports dizziness, drowsiness, head ache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. <PATIENT> was seen at 23:12 minutes by attending physician Dr. <DOCTOR> and was scheduled for emergency head and neck CT with further neurological assessment. At 23:18 he was neurologically assessed by Dr. <DOCTOR> and was HD stable with normal vital signs and therefore and transferred (ID num <IDNUM>) for further radiological investigations.

Result Window

Detected language: en

Masked Text | **Obfuscated Text**

Sacred Heart Memorial

---

1 Verney Drive, 09323, Schroederport, Delaware, Macedonia
Email: Bethuel@hotmail.com,
Phone: (557) 322-0254.

TSICU MRN# 2706237 on 07/08/2019 by ambulance VIN: 6EGBT51VOHY073710.

Bernadine Briar is a 65 y.o. patient admitted to ICU after an MVA on 1905 Highway 97 East, at 23:00 hours. He works as a Paediatric nurse, and long hours of work reported. He reports dizziness, drowsiness, head ache in the frontotemporal region with skin lacerations on his right occipital auricular area.
Mr. Bernadine Briar was seen at 23:12 minutes by attending physician Dr. Dane Dung and was scheduled for emergency head and neck California with further neurological assessment. At 23:18 he was neurologically assessed by Dr. Sabra Cramp and was HD stable with normal vital signs and therefore and transferred (ID num 184378) for further radiological investigations.

Deidentification Tool

TEST DATA

Test your data by applying the following available De-identification tools.

Free Text >>
De-identify free text documents

Table >>
Database, XML, JSON, XLSX, CSV

Documents >>
PDF / DOCX / PPTX / JPEG

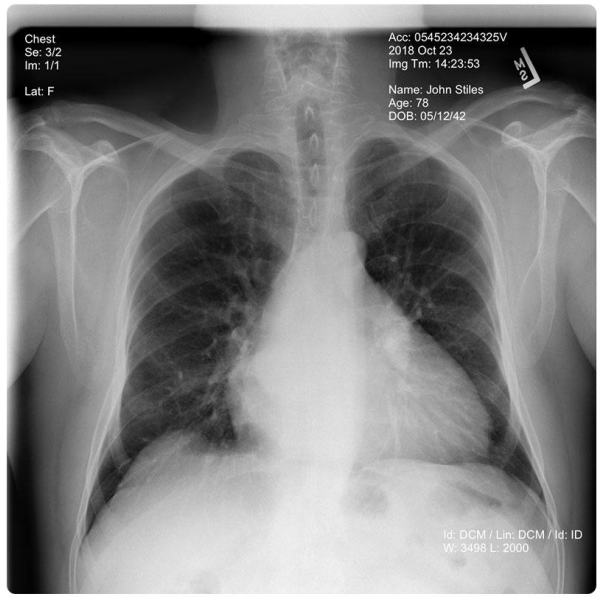
DICOM >>
De-identify DICOM documents

SELECT LANGUAGE

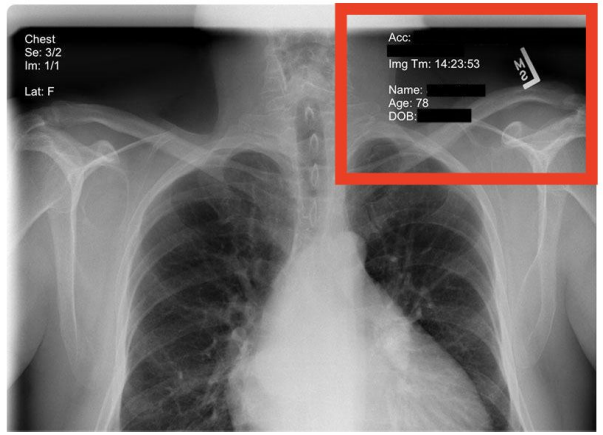
Auto

Powered by John Snow LABS

DICOM Image



De-identified DICOM Image



TEST DATA

Test your data by applying the following available De-identification tools.

Free Text >>
De-identify free text documents

Table >>
Database, XML, JSON, XLSX, CSV

Documents >>
PDF / DOCX / PPTX / JPEG

DICOM >>
De-identify DICOM documents

SELECT LANGUAGE

Auto

Powered by John Snow LABS

Original Data

Case Study Adapted from Addressing HIV Care and Transgender Communities

Mary, a 33 year old Native America male to female transgender person, wants to look into services that you provide. She has not legally changed her name so her documents display her given male name Mark. She is new in transition, dresses in high heels and tight skirts. She produces facial hair (which is exposed). She looks to be very nervous, shy and does not look anyone in the eyes. Mary has been diagnosed HIV positive for three years and has just begun a relationship with Robert who is HIV negative.

Discussion Questions

1. How can the provider establish a sense of trust with the patient?
2. What can the provider do to gain a sense of Mary's knowledge about HIV, how it is contracted, and how it is treated?
3. What social support, if any, should Mary be offered?
4. How can the provider fulfill his/her ethical responsibility to Mary and her partner(s)?
5. Discuss other Cultural Competence issues that may impact retention into care and treatment

Masked Result

Case Study Adapted from Addressing HIV Care and Transgender Communities

[redacted], a 33 year old [redacted] male to female transgender person, wants to look into services that you provide. She has not legally changed her name so her documents display her given male name [redacted]. She is new in transition, dresses in high heels and tight skirts. She produces facial hair (which is exposed). She looks to be very nervous, shy and does not look anyone in the eyes. [redacted] has been diagnosed HIV positive for three years and has just begun a relationship with [redacted] who is HIV negative.

Discussion Questions

1. How can the provider establish a sense of trust with the patient?
2. What can the provider do to gain a sense of [redacted] knowledge about HIV, how it is contracted, and how it is treated?
3. What social support, if any, should [redacted] be offered?
4. How can the provider fulfill his/her ethical responsibility to [redacted] and her partner(s)?
5. Discuss other Cultural Competence issues that may impact retention into care and treatment

Let’s code!