Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets

Veysel Kocaman*, Hasham Ul Haq, David Talby

John Snow Labs Inc., Delaware, USA

* veysel@johnsnowlabs.com

> live at https://deid.johnsnowlabs.com

RWD143

Research Question

- **Clinical Text De-Identification for unstructured clinical notes at scale with state-of-the-art results on benchmark datasets.**
  - While there are models for identification of PHI in text, little work is done to employ ML model for De-Identification at scale.
  - How to scale ML models to process millions of documents without requiring specialized hardware?
  - ***Scale state-of-the-art models to perform de-identification on large datasets.***

Extended Study

- **Expanding ML model to support 7 different languages while maintaining accuracy**
- **Integrating support for rule-based identifiers in the same pipeline.**
- **Developing strategies for redacting information from unstructured text.**
  - **Masking**: Redacting PHI by replacing with fixed values.
  - **Obfuscation:** Redacting PHI by replacing with surrogate values. These surrogate values are generated using **faker**; A module that can generate random names, addresses etc.
- **Comparison with existing solutions from major API providers i.e AWS, GCP, Azure as well as ChatGPT - Spark NLP significantly outperforms competition.**

Table 1: Comparison of the de-identification pipeline with AWS Medical Comprehend (AMC), Microsoft Azure Text Analytics for Health (Azure), & Google Cloud Platform (GCP) Healthcare API on a sample of 100 notes.

		Ours			AMC			Azure			GCP		
Entity	Sample	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Age	95	1.000	1.000	1.000	0.989	0.936	0.962	0.882	0.976	0.927	0.888	0.929	0.908
Date	953	0.999	0.995	0.997	1.000	0.990	0.995	1.000	0.811	0.896	1.000	0.928	0.962
Doctor	402	0.987	0.969	0.978	1.000	0.918	0.957	0.987	0.551	0.707	0.503	0.749	0.602
Hospital	182	0.922	0.911	0.917	0.980	0.810	0.887	0.962	0.573	0.718	0.634	0.829	0.718
Location	47	0.905	0.884	0.894	0.842	0.780	0.810	1.000	0.766	0.867	0.614	0.900	0.730
Patient	115	1.000	0.930	0.964	1.000	0.904	0.950	0.949	0.667	0.783	0.545	0.424	0.477
Phone	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.667	0.800	1.000	0.933	0.966
ID	465	0.941	0.910	0.925	0.922	0.933	0.927	-	-	-	-	-	-
Macro-Avg.				**0.959**			**0.936**			**0.712**			**0.670**
Micro-Avg.				**0.969**			**0.959**			**0.715**			**0.638**

Original Text

Harbor Hospital
---
36 Park Avenue, 95108, San Diego, CA, USA
Email: medunites@harborhospital.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, headache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. 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.

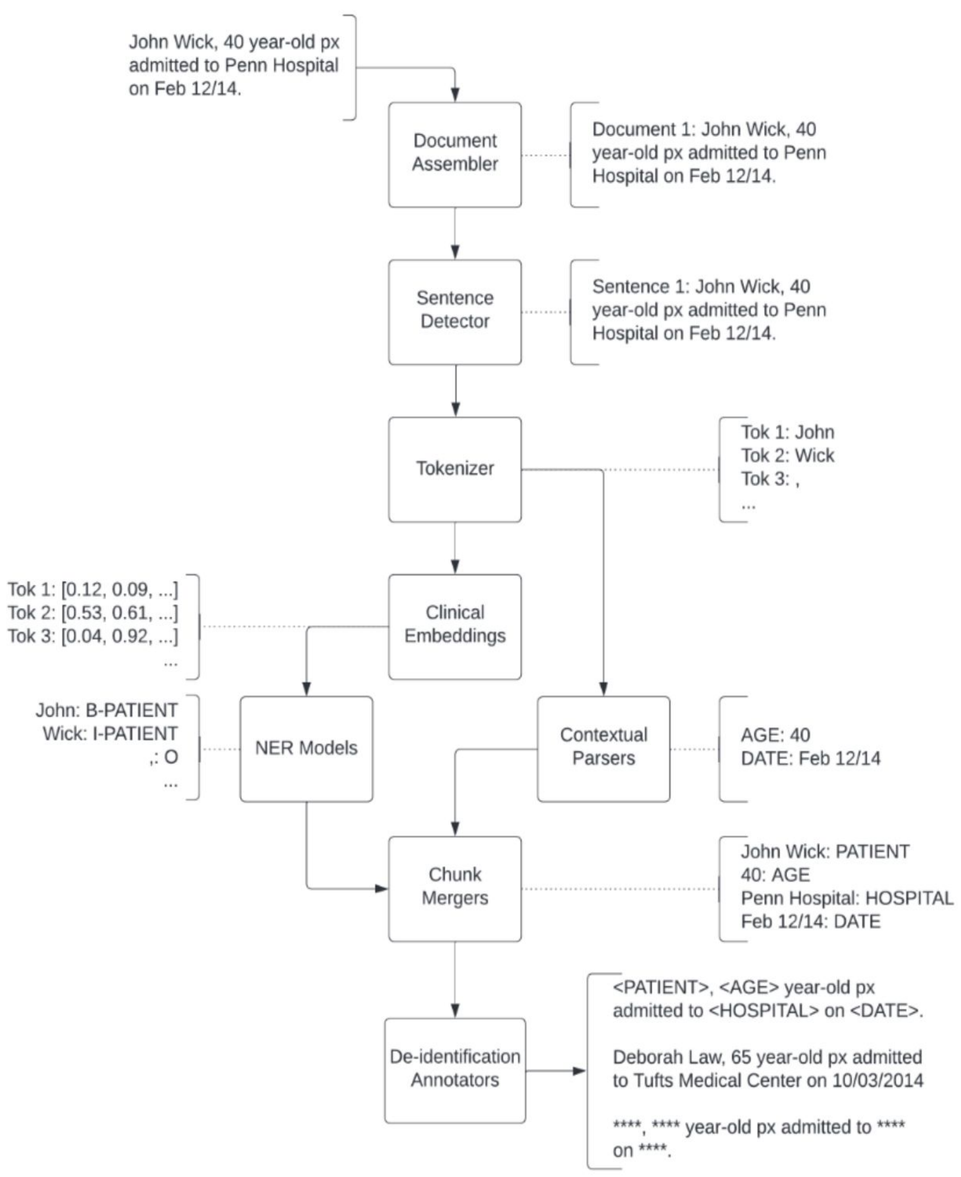
Masked

<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, headache 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.

Obfuscated

MERCY HOSPITAL ARDMORE
---
474 North Yellow Springs Street, 14235, Salt Lake City, Utah, US
Email: dalton@mercyhospital.com,
Phone: (765) 896 92 86.
TSICU MRN# US:3025146 on 15/08/2019 by ambulance VIN: 1AAAA00AAAA111000.
Meldon Lemon is a 58 y.o. patient admitted to ICU after an MVA on 390 40th street, at 23:00 hours.
He works as a special educational needs teacher, and long hours of work reported.
He reports dizziness, drowsiness, headache in the frontotemporal region with skin lacerations on his right occipital auricular area.
Mr. Lemon was seen at 23:12 minutes by attending physician Dr. Evangeline Kelly and was scheduled for emergency head and neck CT with further neurological assessment.
At 23:18 he was neurologically assessed by Dr. Lara Courier and was HD stable with normal vital signs and therefore and transferred (ID num 453267) for further radiological investigations.



Maintaining Data Integrity while performing **Obfuscation** is crucial. The obfuscation module does it by keeping track of similar names and addresses in the text, and replaces them with the same random value. Gender is also considered while selecting the random name. Furthermore, more granular control for shifting dates is provided, so dates can be translated by a certain day offset.

Table 2: Comparison of NER models with full pipelines enriched with regex and contextual parser using Macro-F1 scores.

	English		German		Spanish		Portuguese		Italian		French		Romanian	
Entity	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline
Age	0.910	0.967	0.944	0.965	0.971	0.987	0.963	0.984	0.969	0.984	0.933	0.978	0.840	0.933
Date	0.973	0.988	0.999	0.999	0.965	0.978	0.989	0.995	0.985	0.986	0.991	0.997	0.915	0.952
ID	0.930	0.974	0.974	0.984	0.978	0.994	0.978	0.996	0.980	0.988	0.966	0.983	0.893	0.952
Location	0.803	0.927	0.797	0.855	0.870	0.903	0.958	0.968	0.971	0.985	0.868	0.956	0.596	0.709
Avg.	**0.904**	**0.964**	**0.929**	**0.951**	**0.946**	**0.965**	**0.972**	**0.986**	**0.976**	**0.986**	**0.939**	**0.979**	**0.811**	**0.887**
PHI	0.948	0.982	0.958	0.966	0.974	0.983	0.992	0.994	0.984	0.992	0.986	0.996	0.930	0.957

Conclusion

- State-of-the art results on benchmark datasets.
- It's a python library that can be integrated into existing codebase easily.
- Better performance compared to commercial APIs.
- Easy to deploy on commodity clusters leveraging Spark; on-prem or using cloud.
- Scalable to process millions of notes without specialized hardware.
- Support for 7 different languages with comparable performance.
- Multiple methods to redact information; Obfuscation can be more desirable as it produces text much closer to the original document while removing PHI.