Realizzazione di un servizio per la deidentificazione automatizzata di documenti sanitari

Tesi di Laurea Triennale

UNIVERSITÀ
DI PARMA

Riccardo Fava - 12 Ottobre 2020

Da dove nasce il bisogno di anonimizzare i dati sanitari?

- L'accelerazione che il percorso di digitalizzazione dei servizi sanitari ha avuto negli ultimi anni ha favorito un considerevole aumento della disponibilità di documenti sanitari in formato elettronico, quali cartelle cliniche, referti specialistici, verbali di pronto soccorso e lettere di dimissione ospedaliera.

- I nuovi sistemi per l'estrazione automatica di informazioni dall'enorme mole di dati sanitari distribuiti migliorano prevenzione, diagnosi, cura delle patologie, indagini epidemiologiche e strategie sanitarie.

- L'ostacolo principale nell'analisi del patrimonio informativo sanitario è però legato alle normative sulla Privacy. Dal 2016 è infatti in vigore il GDPR che ha avuto lo scopo di uniformare le leggi sul trattamento dei dati personali all'interno dell'Unione Europea.

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers	(M) Device identifiers and serial numbers
(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section; and
(K) Certificate/license numbers	

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Strategie di deidentificazione

HIPAA Guidelines

Entità da anonimizzare

01 **NAME**
Nomi e cognomi dei pazienti o di loro parenti

02 **HCP**
Nomi e cognomi del personale clinico

03 **AGE**
Età di pazienti, personale, parenti dei pazienti

04 **DATE**
Ogni riferimento temporale a avvenimenti importanti

05 **LOC**
Luoghi specifici (Città, Indirizzi, Luoghi di residenza, ecc...)

06 **ADDRESS**
Indirizzi specifici (Via, numero civico, ecc...)

07 **FAC**
Strutture generiche (Ospedali, Case residenziali, ecc...)

08 **FAC_NAME**
Nome specifico di una struttura.

09 **REL**

Relazioni familiari di pazienti o del personale (Madre, Figlio, Nipote ecc...)

10 **TAX_CODE**

Codice fiscale di pazienti, personale o parenti dei pazienti

11 **PH**

Numeri di telefono

12 **TIME**

Orari specifici di eventi

13 **EMAIL**

Indirizzi e-mail di pazienti, di personale sanitario o strutture cliniche

14 **URL**

Indirizzi web specifici

15 **ZIP**

CAP: Codice di avviamento postale

16 **PLATE**

Targhe di autovetture sanitarie o di automobili appartenenti a personale/pazienti (Ambulanze, ecc...)

Definizione del problema

Il progetto quindi verte sul privare i documenti dei dati sensibili, individuando e anonimizzando le informazioni personali.

Deidentificare significa appunto eliminare la correlazione tra i dati personali e una determinata persona fisica interessata, rendendo impossibile l'identificazione della stessa.

Come si può affrontare il problema?

- Creazione di un dataset annotato partendo da documenti sanitari
- Algoritmi di Named Entity Recognition (NER)
- Approccio supervisionato
- Modelli di word-embeddings

Fasi principali del progetto

FASE 1

Creazione del dataset annotato con Prodigy

FASE 2

Addestramento di un modello baseline con Spacy

FASE 3

Addestramento di un modello con il pre-train di Spacy e i word-embeddings di FastText

FASE 4

Addestramento di un modello con Flair con i word-embeddings di Flair e FastText

FASE 5

Inserimento del modello migliore in una custom recipe di Prodigy per la visualizzazione delle predizioni

Prodigy

Il dataset annotato è stato creato grazie al tool di annotazione Prodigy.
Prodigy permette di annotare documenti in maniera veloce e intuitiva. E' infatti possibile evidenziare le entità all'interno del testo e salvare le annotazioni all'interno del database di Prodigy. Il dataset annotato può essere poi utilizzato per addestrare un modello di machine learning direttamente in Prodigy, oppure è possibile esportarlo e utilizzarlo direttamente nella libreria Spacy.
Nel nostro progetto, per produrre il dataset, sono stati utilizzati dei referti medici.
Il dataset che abbiamo sviluppato comprende circa 1600 record (referti medici annotati).

Confronto tra i modelli

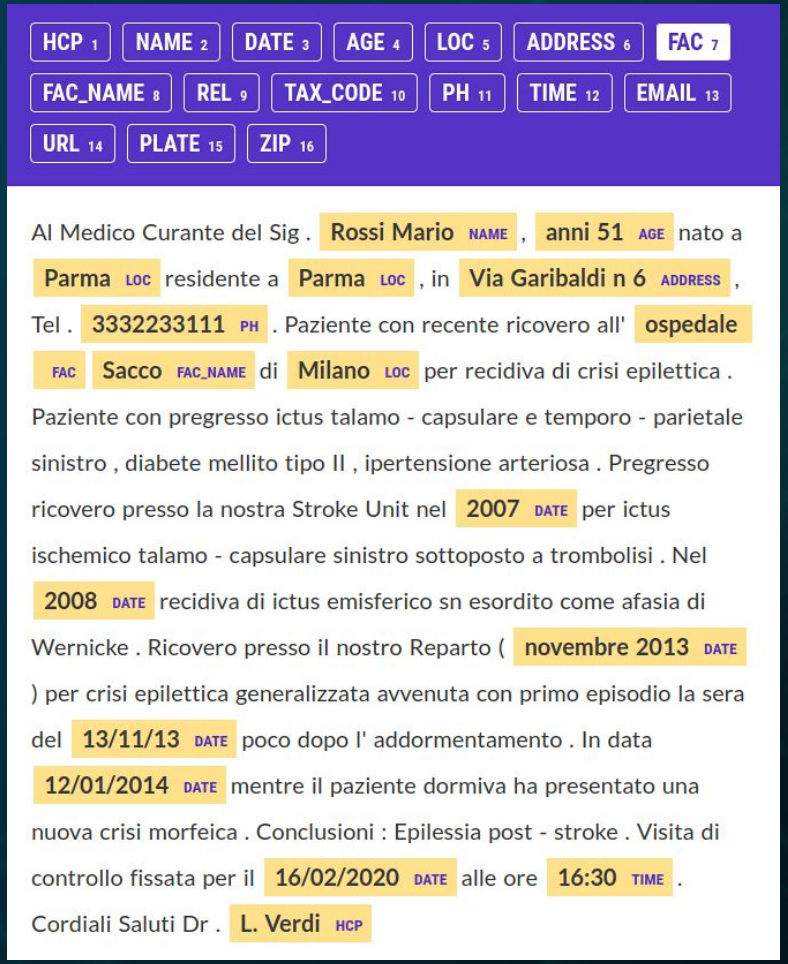
Model	Precision	Recall	F1-Score
Spacy Baseline	0,88	0,88	0,88
Spacy + FastText	0,89	0,89	0,89
Flair + FastText	0.91	0.93	0.92

Confronto con un progetto di anonimizzazione di casi legali francesi

Tag	Precision	Recall	F1-Score
HCP	0.95	0.97	0.96
NAME	0.91	0.87	0.89
AGE	0.90	0.95	0.92
DATE	0.89	0.92	0.90
LOC	0.94	0.92	0.93
FAC	0.87	0.82	0.84
FAC_NAME	0.87	0.87	0.87
REL	0.84	0.91	0.87
ADDRESS	0.90	0.90	0.90
PH	0.99	0.99	0.99
TIME	0.94	0.99	0.96
EMAIL	1.00	1.00	1.00
ZIP	1.00	1.00	1.00

Entity	Precision	Recall	F1
ADDRESS	0.8876	0.8639	0.8756
BAR	1.0000	1.0000	1.0000
COURT	0.9216	0.9495	0.9353
DATE	0.9823	0.9808	0.9815
JUDGE_CLERK	0.9124	0.9605	0.9358
LAWYER	0.9495	0.9617	0.9556
ORGANIZATION	0.9503	0.9324	0.9413
PERS	**0.9570**	**0.9408**	**0.9488**
PHONE_NUMBER	0.9583	0.9200	0.9388
RG	0.9122	0.9353	0.9236

Visualizzazione delle predizioni del modello in Prodigy



Next Steps

- Per migliorare le prestazioni del sistema si possono inserire dei pattern per le entità che ne permettono lo sviluppo (età, date e relazioni...)
- Una volta individuate tutte le occorrenze dei tag bisognerà poi definire una politica di anonimizzazione per ciascuna entità a seconda del caso d'uso nel quale verrà utilizzato questo strumento.
- Il sistema andrà quindi adattato alle politiche di anonimizzazione scelte e poi sarà in grado di deidentificare documenti sanitari di ogni tipo.

THANKS!

RICCARDO FAVA

Un grande ringraziamento al Professor Andrea Prati e al mio tutor aziendale Vieri Emiliani per il grande aiuto e sostegno che mi hanno dato durante lo svolgimento di questo progetto.

UNIVERSITÀ
DI PARMA

MAPS
SHARING KNOWLEDGE