Data output

* 1. **What data should be de-anonymized at the end of the processing pipeline?**

The C2R pipeline

* 1. **What placeholders should be used after NER processing?**
  2. How should the placeholders be stored after NER process?
  3. How should the data be transferred in the process?
  4. Where should the placeholders be stored after NER process and how should they be processed?

Eponymous diseases

1.5 What database should be used to gather a list of eponymous diseases?

1.6 How should eponymous diseases be handled in the NER and placeholder replacement process?

NER models

1.8 How should the models be contained? – huggingface

Introduction

System requirements and challenges

* Implementable into the current text pipeline with sufficient security and deployment options.
* Extreme edgecase handling?
* Lightweight.
  + Limit database hits
* Realtime processing possible
* Both way processing possible 🡪
  + First anonymize
  + Then de-anonymize
* Secure
  + All local, both runner and database.
  + Limited information storage.
  + Active removal of non-needed instances/entities
* Future proof
* Model interchangeable
* Dynamical input and output options
* Storing entities that should be de-anonymized. Rest can be deleted/should not be retained.
* DataStream?
* Classification of Entities into categories
  + Client - data
  + Unknown Entity - data
  + Eponymous disease / medical terms - data
  + Caregiver – data

Literature review

Results

Conclusion

Discussion

What problem are we solving?

Care2Report aims to utilize different types of automated detection mechanisms such as speech recognition and video action recognition in order to automate the process of medical reporting for medical professionals. The processed data is not anonymous which could pose a vulnerability and introduce privacy concerns among its users and therefor endanger the adoptability of the system (<https://link.springer.com/content/pdf/10.1007%2Fs10916-013-9966-z.pdf>). By law (GDPR) it is also required to minimize the amount of security vulnerabilities regarding special category data such as healthcare data (<https://gdpr-info.eu/art-9-gdpr/>).

The GDPR states that special categorical data such as “data concerning health” needs to be processed safely and securely. To minimize leaks and vulnerabilities, data should be processed anonymously. This will also allow to further process the data as anonymous data does not fall under the “special category data” which is described in the GDPR. It is actually completely outside of the scope of EU data protection laws. Which allows for far more extensive processing and use.

***“****1The principles of data protection should apply to any information concerning an identified or identifiable natural person. 2Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. 3To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. 4To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. 5The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. 6This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes****.”***

(GDPR art.26)

Therefor, data anonymization needs to be applied throughout different steps of the process. Starting right after data transcription.

The Care2Report processing pipeline for speech data is as follows:

In this process all steps are performed on personal data. This means a greater risk for privacy leaks and vulnerabilities. To minimize data exposure, anonymization can be applied. Using the method described in this paper, the amount of steps processing non-anonymized data will be reduced to two out of five. The result will be as follows:

Important factors:

* Placeholder output 🡪 standardized? Impact on actual summarization.
* Delivery and output methods 🡪 impact on speed.
* Handling of undetected Names Entities 🡪 can be ignored, because context is sufficiently changed?
* Eponymous disease

Why is it necessary to solve this issue?

How are we going to solve this issue?

Technologies user:

* Postgres database
* Python
* Huggingface model repository
* Snomed CT disease database
* C2R pipeline

How should audio transcription data be anonymized before and after automated summarization to resolve privacy concerns regarding data processing?

Sessionid assigned

|

Language detection

|

NER processing using preset models (huggingface)

|

Placeholder replacement

|

Output cleaned text

|

…..

|

Input sentences from processing tasks

|

Check for placeholders in output and de-anonymize

|

Output de-anonymized text