REPORT: Group H

I. Introduction

i. Background

In the digital age, data is one of the most important assets for every business. As such, the care of this asset must be a primary concern. Data governance has become an important field to achieve a purposeful, responsible, legal and efficient treatment of data, in which control over such data is fundamental.[1]

A proper control of data throughout its lifecycle can result in saving and preventing costs and risks, both from an operational[2] and legal perspective. In this particular, the period of time data is retained by a business in the European Union (EU) is of utmost importance for the compliance with legal provisions such as the General Data Protection Regulation (GDPR), which establishes a limit to the storage of data that may allow the identification of individuals (personal data) for longer than is necessary, for the purposes for which the personal data are processed (storage limitation principle).[3] Likewise, the GDPR requires that personal data be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimisation principle).[4]

To comply with such principles, businesses may resort to the deletion of the data, however, some of that data may still be useful to perform other tasks and thus, recycling that data is a way to save costs by optimizing its potential.

ii. Our project

Our project aims to provide organizations and individuals in the medical field with the ability to maximize the value of the data they collect and process, specifically in relation to their legal obligations related to personal data, by providing a tool for data anonymization powered by Artificial Intelligence (AI). With it, they will be able to "clean" their data by extracting those data points that might be used to identify a person from the rest of the data that could be further reused for other purposes. In addition, by having a better control and understanding of the time of retention of their data and its uses, they may also reduce their expenses regarding storage and archiving.[5]

[1] Marijn Janssen and others, 'Data Governance: Organizing Data for Trustworthy Artificial Intelligence' (2020) 37 Government Information Quarterly 101493, 2.

[2] Vijay Khatri and Carol V Brown, 'Designing Data Governance' (2010) 53 Communications of the ACM 148, 151.

[3] European Parliament and Council Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) [2016] ('GDPR'), art 5, (e).

[4] Ibid, (c).

[5] Khatri and Brown (n 2) 151.

This report presents the activities carried out for the design of our AI tool and how it overcomes some limitations of previous tools with similar features. We further explain the practicalities and advantages of our work and the developments we hope to achieve in the future.

With AI, we seek to support a better data governance design through more robust control and management of data. Our tool not only serves as an instrument for compliance with legal obligations in relation to personal data but also allows organizations and individuals obtain the maximum potential of the data they collect and process.[6]

Although this project focuses on the medical field, the ultimate goal of our work is to serve as a reference in other specific contexts for individuals and organizations.

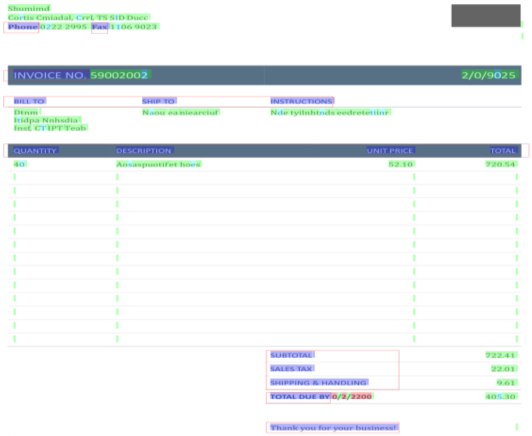
II. Previous Work

i. Sypht-team/pdf-anonymizer: Sypht-team/pdf-anonymizer:

Developed in JavaScript, this tool anonymizes PDF documents by first detecting the personal data and then randomizing its characters and the typography and format of the data is preserved. However, the randomization of dates will still have an output of numbers but often the numbers would be absurd, for example: the year 9025. The tool takes the PDF document as an input and produces 2 possible output PDF-document in which:

1. Personal data of the document are randomized.
2. Data sections within document are highlighted, explaining the reasoning of the tool:

- Green: The randomized personal data.
- Violet: The metadata of the personal data.
- Blue: Personal data could not be randomized due to being within white-listed zones.
- Red: Personal data could not be randomized due to the lack of characters in the same font.
- Orange: Personal data that were not substituted and are not punctuated.



Shumimd
Cortis Cmiadal, Crrl, TS SID Ducc
Phone 0222 2995 Fax 1106 9023

INVOICE NO. 59002002 2/0/9025

BILL TO	SHIP TO	INSTRUCTIONS
Dtnm Itidpa Nnhsdia Insf, CT IPT Teab	Naou ea niearciuf	Nde tyilnhtnds eedretetiinr

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
40	Aosaspuotifet hoes	52.10	720.54

SUBTOTAL	722.41
SALES TAX	22.01
SHIPPING & HANDLING	9.61
TOTAL DUE BY 0/2/2200	405.30

Thank you for your business!

Limitations:

- No explanation behind the data anonymization.
- The regeneration of the PDF with randomized data using different typography (shape, font style, height, and weight) causes often problems

[6] Simson L Garfinkel, 'De-Identification of Personal Information' (National Institute of Standards and Technology 2015) NIST IR 8053 4 <https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf> accessed 3 March 2024.

- It is very difficult to perform analytics with randomized data thus losing one of the main advantages of anonymizing data in the first place.

ii. ArtLabss/open-data-anonymizer:

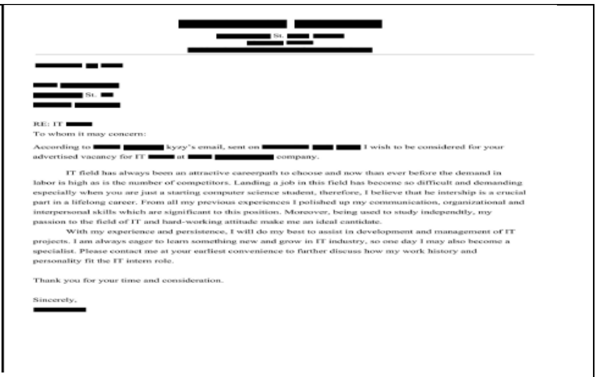
Unlike the previous tool, this one is not only limited to PDF files, it has also the option of both tabular data along with images and facial anonymization techniques. Written in Python, the tool relies on regular expressions to identify the personal data from a PDF document, it also uses techniques of Optical character recognition to mask the personal data after its identification.

The tool gives the option to decide which personal data could be masked, for example emails only. The data that is considered personal data is then masked later:

<table>
<tr>
<td>

- Names
- Dates
- Location
- Job status
- Companies Name

</td>
<td>



[redacted]
[redacted] St. [redacted]
[redacted]
[redacted]

[redacted]

[redacted]
[redacted] St. [redacted]
[redacted]

RE: IT [redacted]
To whom it may concern:

According to [redacted] kyzy's email, sent on [redacted] I wish to be considered for your advertised vacancy for IT [redacted] at [redacted] company.

IT field has always been an attractive careerpath to choose and now than ever before the demand in labor is high as is the number of competitors. Landing a job in this field has become so difficult and demanding especially when you are just a starting computer science student, therefore, I believe that he intership is a crucial part in a lifelong career. From all my previous experiences I polished up my communication, organizational and interpersonal skills which are significant to this position. Moreover, being used to study independtly, my passion to the field of IT and hard-working attitude make me an ideal cantidate.

With my experience and persistence, I will do my best to assist in development and management of IT projects. I am always eager to learn something new and grow in IT industry, so one day I may also become a specialist. Please contact me at your earliest convenience to further discuss how my work history and personality fit the IT intern role.

Thank you for your time and consideration.

Sincerely,
[redacted]

</td>
</tr>
</table>

Limitations:

- Regular expressions will produce poor output of personal data if different example formats are introduced.
- No explanation behind the data anonymization.
- Only the Personal data on header and the introduction are masked, but the other sections are not masked.

III. Activities and Outputs

i. Data Construction

We used synthetic data for our study, generating a model of medical reports through ChatGPT-3.5, tailored to mimic those produced by healthcare professionals. Subsequently, we introduced variations in the personal data sections. We manually **created approximately 100 medical reports** to supply our natural language processor with diverse training examples.

ii. Medical reports

We relied on a variety of personal medical reports, which contain a lot of information about a medical consultation; patient's history, examinations performed on the patient, the diagnosis or recommendations to be followed, etc. We focused our project on the following categories of personal data:

- Person: The **name** of the patient and the doctor concerned.
- Geo-Political Entity (GPE): The **location** of the patient's home and the hospital.
- Date: The **dates** of the patient's birth and of the medical consultation.
- ID: The patient's **ID number**.
- Insurance: The patient's **insurance number**.
- Email: The patient's **email address**.
- Credit card: The patient's **credit card number**.

iii. Legislation involved

We based our data categorization on the definition of personal data provided by the GDPR, given that our tool is meant to be used initially in the EU.

Article 4 of the GDPR defines personal data as "Any information relating to an identified or identifiable natural person, who can be recognized directly or indirectly, particularly by reference to an identifier."

According to such definition, each of the chosen categories contain personal data, as explained below:

- **Person**: Based on Article 4(1) of the GDPR, this information contains data that can be related to an identified or identifiable natural person, in this case directly by reference to a name. The name and surname of an individual are the most common identifiers.
- **Geo-Political Entity (GPE):** The GPE category includes information indirectly or directly linked to a person, notably through location data. The personal address of an individual constitutes information that can be used to directly identify an individual. The address of the hospital is an identifier that, when combined with other pieces of information, can be associated with a person.
- **Date**: The date of birth of a person and the date of a medical consultation contains data that, combined with other pieces of information, such as a name, will allow the individual to be identified.
- **ID**: The ID in the medical report contains information that can be related to an identified or identifiable natural person, in this case by reference to an identification number. An identification number is a unique identifier assigned to each patient to distinguish them within the healthcare system. This ID number serves as a reference point for accessing and managing an individual's medical information. Therefore, an identification number contains data that can be used to directly identify a person.

- **Insurance**: This specific information contains data that can be related to an identified or identifiable natural person (directly or indirectly). The insurance number of an individual is considered personal data due to its direct association with an individual's financial and health-related information. This number uniquely identifies an individual.
- **Email**: An email address is considered personal data because it directly relates to an individual by reference to an online identifier.
- **Credit Card:** A credit card is considered an identifier that expresses the economic identity of a natural person and that can be directly or indirectly related to an individual.

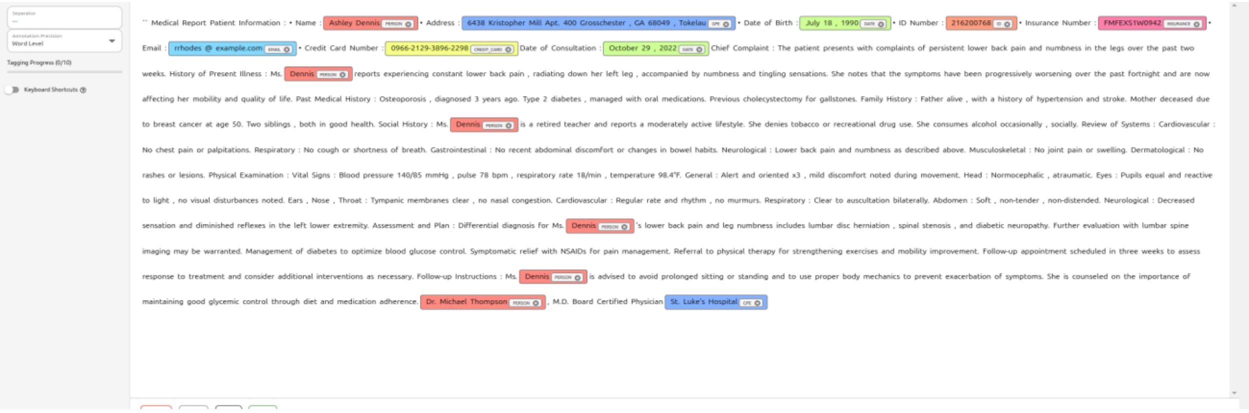
iv. Data Annotation

The process of data annotation was facilitated by utilizing the Ner-annotator website. Within this phase, seven distinct categories of personal data were identified and annotated within the documents, as shown in the image below.

1. Person (Name & Surname).
2. Geo-Political Entity (GPE).
3. Date (birth & Consultation dates).
4. ID.
5. Insurance number.
6. Emails.
7. Credit Cards.

The annotator operates by taking a text file as input and generating a JSON file as output. Following this, a minor preprocessing step is applied to the JSON file before it is utilized as input for the selected Natural Language Processor (NLP).

v. Artificial Intelligence



`` Medical Report Patient Information : • Name : Ashley Dennis PERSON • Address : 6438 Kristopher Mill Apt. 400 Grosschester , GA 68049 , Tokelau GPE • Date of Birth : July 18 , 1990 DATE • ID Number : 216200768 ID • Insurance Number : FMFEXS1W0942 INSURANCE • Email : rrhodes @ example.com EMAIL • Credit Card Number : 0966-2129-3896-2298 CREDIT_CARD Date of Consultation : October 29 , 2022 DATE Chief Complaint : The patient presents with complaints of persistent lower back pain and numbness in the legs over the past two weeks. History of Present Illness : Ms. Dennis PERSON reports experiencing constant lower back pain , radiating down her left leg , accompanied by numbness and tingling sensations. She notes that the symptoms have been progressively worsening over the past fortnight and are now affecting her mobility and quality of life. Past Medical History : Osteoporosis , diagnosed 3 years ago. Type 2 diabetes , managed with oral medications. Previous cholecystectomy for gallstones. Family History : Father alive , with a history of hypertension and stroke. Mother deceased due to breast cancer at age 50. Two siblings , both in good health. Social History : Ms. Dennis PERSON is a retired teacher and reports a moderately active lifestyle. She denies tobacco or recreational drug use. She consumes alcohol occasionally , socially. Review of Systems : Cardiovascular : No chest pain or palpitations. Respiratory : No cough or shortness of breath. Gastrointestinal : No recent abdominal discomfort or changes in bowel habits. Neurological : Lower back pain and numbness as described above. Musculoskeletal : No joint pain or swelling. Dermatological : No rashes or lesions. Physical Examination : Vital Signs : Blood pressure 140/85 mmHg , pulse 78 bpm , respiratory rate 18/min , temperature 98.4°F. General : Alert and oriented x3 , mild discomfort noted during movement. Head : Normocephalic , atraumatic. Eyes : Pupils equal and reactive to light , no visual disturbances noted. Ears , Nose , Throat : Tympanic membranes clear , no nasal congestion. Cardiovascular : Regular rate and rhythm , no murmurs. Respiratory : Clear to auscultation bilaterally. Abdomen : Soft , non-tender , non-distended. Neurological : Decreased sensation and diminished reflexes in the left lower extremity. Assessment and Plan : Differential diagnosis for Ms. Dennis PERSON 's lower back pain and leg numbness includes lumbar disc herniation , spinal stenosis , and diabetic neuropathy. Further evaluation with lumbar spine imaging may be warranted. Management of diabetes to optimize blood glucose control. Symptomatic relief with NSAIDs for pain management. Referral to physical therapy for strengthening exercises and mobility improvement. Follow-up appointment scheduled in three weeks to assess response to treatment and consider additional interventions as necessary. Follow-up Instructions : Ms. Dennis PERSON is advised to avoid prolonged sitting or standing and to use proper body mechanics to prevent exacerbation of symptoms. She is counseled on the importance of maintaining good glycemic control through diet and medication adherence. Dr. Michael Thompson PERSON , M.D. Board Certified Physician St. Luke's Hospital GPE

The model utilized for this task was SpaCy which is an open-source library for NLP in Python. It is designed to be efficient, fast and capable of handling large volumes of text. SpaCy provides tools and pre-trained models for various NLP tasks such as tokenization and named entity recognition (NER) which is what is needed for this project. The idea was mainly to create our synthetic data, define our categories of personal data, and then fine-tune one of SpaCy NER models to detect the personal data within our documents.

Following the adjustment of the machine setup to utilize Graphics Processing Unit (GPU) resources instead of Central Processing Unit (CPU), a notable enhancement in training speed was observed, resulting in a threefold increase. This optimization led to an overall reduction in training time to just 5 minutes 54 seconds. The utilized machine boasted the following characteristics:

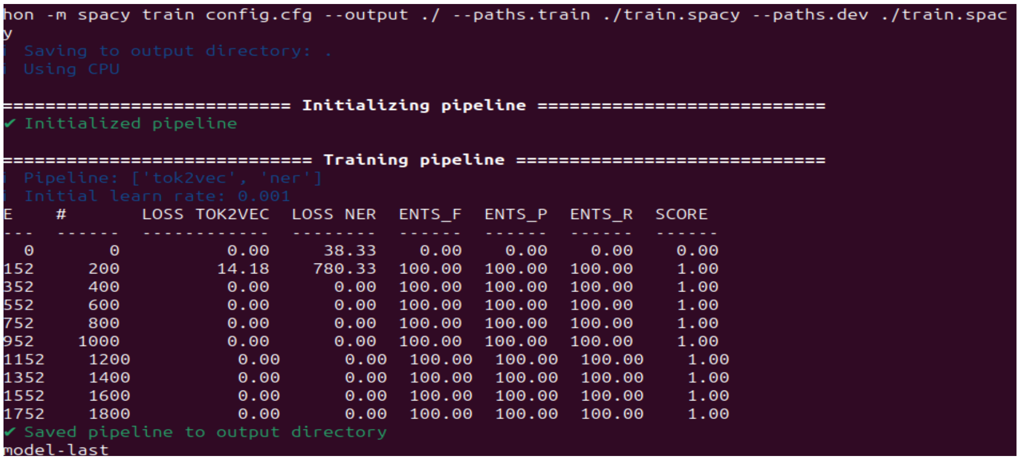
- Ram: 16gbs
- CPU: AMD Ryzen 7 6800H
- GPU: RTX 3050 TI

We have utilized the SpaCy model named “en_core_web_trf” based on the positive results given compared to other models. The relevant parameters used during the fine-tuning were:

- Vectors: " en_core_web_trf"
- batch_size: 1000
- eval_frequency: 200
- patience: 1600
- learn_rate: 0.001
- Dropout: 0.1

In the following image, certain metrics are listed as shown below:

- Loss TOK2VEC & NER: The mean squared error (MSE) in the context of Tok2Vec & NER training.
- Score: Accuracy which is (TP+TN)/total Predictions.
- Ents_f: Weight for the F1 score of entities.
- Ents_p: Weight for the precision score of entities.
- Ents_r: Weight for the recall score of entities.
- Ents_per_type: Weight for entity scores per type.



```
hon -m spacy train config.cfg --output ./ --paths.train ./train.spacy --paths.dev ./train.spac
y
  Saving to output directory: .
  Using CPU

========================== Initializing pipeline ===========================
✔ Initialized pipeline

============================ Training pipeline =============================
  Pipeline: ['tok2vec', 'ner']
  Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     38.33    0.00    0.00    0.00    0.00
152     200         14.18    780.33  100.00  100.00  100.00    1.00
352     400          0.00      0.00  100.00  100.00  100.00    1.00
552     600          0.00      0.00  100.00  100.00  100.00    1.00
752     800          0.00      0.00  100.00  100.00  100.00    1.00
952    1000          0.00      0.00  100.00  100.00  100.00    1.00
1152    1200          0.00      0.00  100.00  100.00  100.00    1.00
1352    1400          0.00      0.00  100.00  100.00  100.00    1.00
1552    1600          0.00      0.00  100.00  100.00  100.00    1.00
1752    1800          0.00      0.00  100.00  100.00  100.00    1.00
✔ Saved pipeline to output directory
model-last
```

-SpaCY Training Evaluation-

The picture shows different evaluation metrics and their values during each epoch, stopping at 1800 saving two models, best and last.

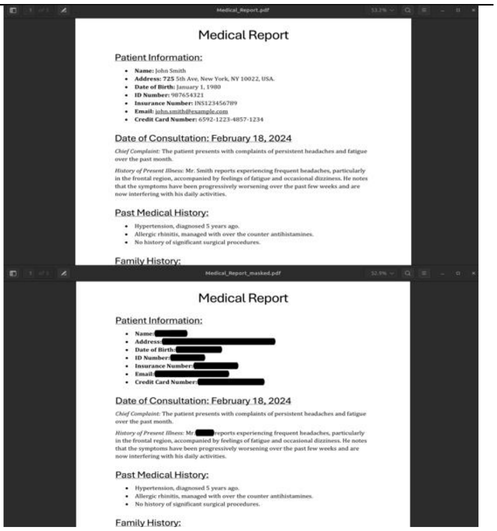
vi. Output

The AI system we developed can perform two distinct tasks: anonymization and personal data identification.

1. For anonymization, the system masks the detected data by overlaying a black box, effectively obscuring sensitive information.
2. In the case of personal data identification, the system **highlights** the detected personal data and provides explanations regarding why it was classified as such. This classification is reinforced by references to relevant articles of the GDPR, supported by additional clarifications as needed.

The Data Anonymizer employs a systematic procedure to execute the data masking process, which includes the following steps:

1. Utilizing the Python library pymuPDF to scan the PDF-document comprehensively.
2. Extracting all words present within the document, thereby obtaining the complete textual content.
3. Feeding the entirety of the text into the AI model, which generates annotations for each identified entity (Text section that is considered personal data).
4. Locating the extracted entities within the original PDF-document and subsequently applying masking techniques to conceal this sensitive information effectively.



Medical Report

Patient Information:

- **Name:** John Smith
- **Address: 725** 5th Ave, New York, NY 10022, USA.
- **Date of Birth:** January 1, 1980
- **ID Number:** 987654321
- **Insurance Number:** INS123456789
- **Email:** john.smith@example.com
- **Credit Card Number:** 6592-1223-4857-1234

Date of Consultation: February 18, 2024

Chief Complaint: The patient presents with complaints of persistent headaches and fatigue over the past month.

History of Present Illness: Mr. Smith reports experiencing frequent headaches, particularly in the frontal region, accompanied by feelings of fatigue and occasional dizziness. He notes that the symptoms have been progressively worsening over the past few weeks and are now interfering with his daily activities.

Past Medical History:

- Hypertension, diagnosed 5 years ago.
- Allergic rhinitis, managed with over the counter antihistamines.
- No history of significant surgical procedures.

Family History:

Medical Report

Patient Information:

- **Name:**[redacted]
- **Address:**[redacted]
- **Date of Birth:**[redacted]
- **ID Number:**[redacted]
- **Insurance Number:**[redacted]
- **Email:**[redacted]
- **Credit Card Number:**[redacted]

Date of Consultation: February 18, 2024

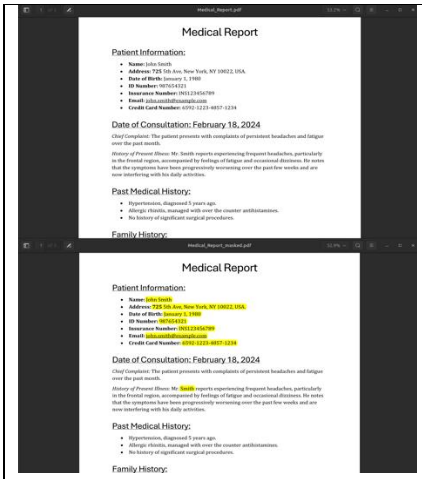
Chief Complaint: The patient presents with complaints of persistent headaches and fatigue over the past month.

History of Present Illness: Mr.[redacted]reports experiencing frequent headaches, particularly in the frontal region, accompanied by feelings of fatigue and occasional dizziness. He notes that the symptoms have been progressively worsening over the past few weeks and are now interfering with his daily activities.

Past Medical History:

- Hypertension, diagnosed 5 years ago.
- Allergic rhinitis, managed with over the counter antihistamines.
- No history of significant surgical procedures.

Family History:

 Medical_Report.pdf Medical Report Patient Information: • **Name:** John Smith • **Address: 725** 5th Ave, New York, NY 10022, USA. • **Date of Birth:** January 1, 1980 • **ID Number:** 987654321 • **Insurance Number:** INS123456789 • **Email:** john.smith@example.com • **Credit Card Number:** 6592-1223-4857-1234 Date of Consultation: February 18, 2024 *Chief Complaint:* The patient presents with complaints of persistent headaches and fatigue over the past month. *History of Present Illness:* Mr. Smith reports experiencing frequent headaches, particularly in the frontal region, accompanied by feelings of fatigue and occasional dizziness. He notes that the symptoms have been progressively worsening over the past few weeks and are now interfering with his daily activities. Past Medical History: • Hypertension, diagnosed 5 years ago. • Allergic rhinitis, managed with over the counter antihistamines. • No history of significant surgical procedures. Family History: Medical_Report_masked.pdf Medical Report Patient Information: • **Name:** John Smith • **Address: 725** 5th Ave, New York, NY 10022, USA. • **Date of Birth:** January 1, 1980 • **ID Number:** 987654321 • **Insurance Number:** INS123456789 • **Email:** john.smith@example.com • **Credit Card Number:** 6592-1223-4857-1234 Date of Consultation: February 18, 2024 *Chief Complaint:* The patient presents with complaints of persistent headaches and fatigue over the past month. *History of Present Illness:* Mr. Smith reports experiencing frequent headaches, particularly in the frontal region, accompanied by feelings of fatigue and occasional dizziness. He notes that the symptoms have been progressively worsening over the past few weeks and are now interfering with his daily activities. Past Medical History: • Hypertension, diagnosed 5 years ago. • Allergic rhinitis, managed with over the counter antihistamines. • No history of significant surgical procedures. Family History:	The Data Highlighter employs a systematic procedure like the Data Anonymizer with slight modifications in the subsequent processes: 1. Steps 1 to 3 remain unchanged: 4. After annotation, the Data Highlighter locates the extracted entities within the original PDF-document. 5. Rather than masking, the Data Highlighter applies highlighting techniques to emphasize the identified entities. Simultaneously, each highlighted entity is accompanied by a specific GDPR explanation detailing why it was classified as personal data.

IV. Discussion

i. Feasibility

Our project offers a **simple, fast, and advantageous solution** for organizations obliged to comply with data minimization and storage limitation requirements outlined in the GDPR. Here's why:

- **Ease of Implementation:** The project leverages existing AI technologies, making it **adaptable to various organizational structures and technical setups**.
- **Minimal Disruption:** The system integrates seamlessly with existing workflows, requiring minimal changes to current data management processes.
- **Cost-Effectiveness:** Compared to traditional anonymization methods, this AI-powered approach offers **significant cost savings** in terms of time, resources, and manpower required.

However, it's important to acknowledge the following considerations:

- **Data Complexity:** The accuracy of the AI model might require adjustments based on the specific data formats, languages, and complexities encountered within an organization's data landscape.
- **Continuous Improvement:** As with any AI system, ongoing monitoring and refinement of the model might be necessary to maintain optimal performance and address evolving data privacy regulations.

Overall, the project demonstrates a highly feasible and advantageous approach not only to data anonymization, but also to storage limitation offering organizations a simple and effective solution for achieving compliance while preserving valuable non-personal data.

- **Reduces Data Retention Needs:** By identifying and masking personal data within files, our project allows organizations to **retain non-personal data** for further use even after fulfilling the initial purpose for collecting the data. This eliminates the need to delete the entire file, adhering to the storage limitation principle while still allowing organizations to utilize valuable insights from the non-personal data.

- **Minimizes Data Subject Risk:** By removing or obscuring personal data, our project **reduces the amount of sensitive information** organizations need to store. This lowers the potential risk of data breaches and minimizes the impact on individuals if a breach occurs, aligning with the GDPR's objective of protecting data subjects.
- **Enables Secure Data Exchange:** Anonymized data allows for **safer and easier sharing** within the organization and potentially with external entities. This facilitates collaboration and knowledge sharing while minimizing privacy concerns associated with exchanging personal data.
- **Simplifies Data Management:** our project can help organizations **streamline their data management processes** by eliminating the need to determine which files require complete deletion due to personal data content. This reduces the administrative burden and simplifies compliance efforts.

ii. Practicality

Building upon the project's foundation, several areas can benefit from further development to enhance its practicality:

- **Scalability and Performance:** Optimizing the system to handle large data volumes efficiently is crucial for broader organizational adoption.
- **User Interface and Training:** Developing intuitive contextual interfaces and comprehensive training materials will ensure user-friendliness and facilitate seamless integration into existing workflows.
- **Security Enhancements:** Implementing robust security measures throughout the anonymization process is essential to safeguard sensitive data and maintain user trust.

By addressing these practical considerations, we can further strengthen the project's real-world applicability and empower organizations to confidently leverage anonymized data for various purposes while adhering to data minimization and storage limitation principles. Continuous evaluation, improvement, and adaptation are key to ensuring the long-term success and widespread adoption of our solution.

V. Conclusion

Through this report, we have presented an AI model capable of performing extremely useful tasks that allow better control over the data organizations and individuals collect and process in the medical field. By providing a useful and fast tool capable of identifying and anonymizing personal data, we expect that our work will serve as a practical detailed guideline to realize the same data minimization work in other specific contexts for individuals and organizations to have a tighter control and management over their data, while facilitating the compliance with data protection legislation.

Future work

As AI models are in continuous development and monitoring, we acknowledge the challenges we must address for a broader use of our tool in different data landscapes. Accordingly, we intend to improve our model by creating an effortless and practical interface for users, as well as enriching the data training sets for the model, to enhance its output capabilities so that it can be used in a more general data landscape.