HSLU Hochschule Luzern

Hochschule Luzern (Schweiz)
– Departement Informatik –

Entwicklung eines funktionsfähigen Webportals zur automatisierten und dynamischen Datenanonymisierung

Bachelorarbeit

vorgelegt am Departement Informatik der Hochschule Luzern (Schweiz) in Anbetracht zur Erreichung des akademischen Grades Bachelor in Informatik.

von

Pascal Derungs
Matrikelnummer: 22-897-904

betreut von

Radwan Eskhita

Bachelorarbeit an der Hochschule Luzern – Informatik

Titel: Entwicklung eines funktionsfähigen Webportals zur automatisierten und dynamischen Datenanonymisierung

Student: Pascal Derungs

Studiengang: Bachelor of Science (BSc) Informatik

Abschlussjahr: 2025

Betreuer: Radwan Eskhita

Experte: Dominique Portmann

Auftraggeber: HSLU I

Codierung / Klassifizierung der Arbeit:

☒ Öffentlich (Normalfall)

☐ Vertraulich

Eidesstattliche Erklärung
Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig und ohne unerlaubte fremde Hilfe angefertigt habe, alle verwendeten Quellen, Literatur und andere Hilfsmittel angegeben habe, wörtlich oder inhaltlich entnommene Stellen als solche kenntlich gemacht habe das Vertraulichkeitsinteresse des Auftraggebers wahren und die Urheberrechtsbestimmungen der Hochschule Luzern respektieren werde.

Ort / Datum, Unterschrift: _____

Abgabe der Arbeit auf der Portfolio Datenbank
Bestätigungsvisum Studentin/Student Ich bestätige, dass ich die Bachelorarbeit korrekt gemäss Merkblatt auf der Portfolio Datenbank ablege. Die Verantwortlichkeit sowie die Berechtigungen gebe ich ab, so dass ich keine Änderungen mehr vornehmen oder weitere Dateien hochladen kann.

Ort / Datum, Unterschrift: _____

Abstract

Die zunehmende Digitalisierung führt zu einem exponentiellen Anstieg personenbezogener Daten und bringt erhebliche Risiken für die Privatsphäre mit sich. Selbst anonymisierte Daten können durch moderne Analyseverfahren potentiell re-identifiziert werden. Ziel dieser Arbeit ist die Entwicklung eines Webportals zur automatisierten und dynamischen Datenanonymisierung. Das Webportal kombiniert zur Erkennung sensibler Daten heuristische Verfahren wie benutzerdefinierte Labels und reguläre Ausdrücke mit Natural Language Processing (NLP)-basierter Named Entity Recognition (NER) auf Basis von spaCy und Microsoft Presidio. Für die Anonymisierung kommen wissenschaftlich fundierte Modelle wie k-Anonymität, ℓ-Diversität und t-Closeness zum Einsatz. Integriert und evaluiert werden diese Modelle über Bibliotheken wie Anjana und pyCANON. Die Architektur des Prototyps basiert auf Python und Streamlit, wobei die Modularität und Nachvollziehbarkeit der Anonymisierung im Vordergrund stehen. Die Ergebnisse dieser Arbeit zeigen, dass sensible Informationen automatisch erkannt und durch individuell konfigurierbare Anonymisierungsmodelle modifiziert werden können. Visualisierungen wie Vorher-Nachher-Vergleiche, Entropieanalysen sowie weitere Metriken ermöglichen eine transparente Bewertung der Anonymisierungsauswirkungen. Das entwickelte Webportal stellt eine solide technische Basis dar, um Datenschutzrisiken zu reduzieren. Es bietet Potenzial für zukünftige Erweiterungen, etwa durch Differenzielle Privatsphäre, rechtliche Prüfmechanismen, oder intelligente Modellvorschläge.

Keywords: spaCy, Presidio, pyCANON, Anjana, sensible Daten, Streamlit

Verdankung
Mein besonderer Dank gilt meinem Betreuer Radwan Eskhita, der mich mit ausserordentlichem Engagement, grosser Fachkenntnisse und stets konstruktivem Feedback durch alle Phasen dieser Arbeit begleitet hat. Seine wertvollen Rückmeldungen sowie sein offenes Ohr für Fragen und Ideen haben wesentlich zum Gelingen dieser Bachelorarbeit beigetragen. Ich habe seine professionelle und motivierende Art sehr geschätzt. Ebenso möchte ich mich bei Dominique Portmann bedanken, der als Experte die Arbeit begleitet hat. Auch wenn der Austausch begrenzt war, danke ich ihm für seine Bereitschaft, Zeit und Expertise in die Bewertung dieser Arbeit einzubringen. Abschliessend danke ich allen, die mich während dieser Arbeit unterstützt und ermutigt haben.

Pascal Derungs, 4. Juni 2025

Inhaltsverzeichnis

1 Einleitung

1.1 Problemstellung und Motivation

Die Digitalisierung sowie der wachsende Einsatz des Internets der Dinge und sozialer Netzwerke haben das weltweite Datenvolumen in den letzten Jahrzehnten exponentiell wachsen lassen [1, S. 1]. Gleichzeitig entstehen neue Herausforderungen für den Datenschutz, insbesondere bei der Anonymisierung personenbezogener Daten. Wie Ciriani, De Capitani Di Vimercati, Foresti et al. [2, S. 1] erläutern, erlauben moderne Analysemethoden die Verknüpfung anonymisierter Daten mit anderen Quellen. Solche Verknüpfungsangriffe machen es möglich, vermeintlich anonyme Daten zu re-identifizieren. Dies bringt erhebliche Datenschutzrisiken mit sich.

Daher ist eine flexible Lösung erforderlich, die sowohl die automatisierte Erkennung sensibler Daten als auch eine transparente Visualisierung der Anonymisierungsauswirkungen ermöglicht.

1.2 Ziel der Arbeit

Diese Arbeit befasst sich mit der Entwicklung eines funktionsfähigen Webportals zur automatisierten und dynamischen Datenanonymisierung. Die Lösung soll zunächst sensible Daten innerhalb eines Datensatzes automatisch identifizieren und klassifizieren. Anschliessend erfolgt die flexible Anonymisierung mit Orientierung an den Vorgaben der Datenschutz-Grundverordnung (vgl. Art. 25 DSGVO, [3]) und des revidierten Schweizer Datenschutzgesetzes (vgl. Art. 5 f. revDSG, [4]). Dabei soll eine Visualisierung sowohl der sensiblen Daten als auch der Auswirkungen der Anonymisierung integriert werden. So soll Transparenz über den Grad der Anonymisierung sowie über mögliche Informationsverluste geschaffen werden.

Zur Evaluation wird das System mit realitätsnahen Testdaten aus verschiedenen Anwendungsfeldern, beispielsweise dem Gesundheits- und E-Commerce-Bereich, getestet. Ziel ist eine praxisnahe und flexible Datenschutzlösung, die über spezifische Branchenanwendungen hinaus nutzbar ist.

Aus dieser Zielstellung ergeben sich folgende Forschungsfragen:

RQ1 Wie können sensible Daten in strukturierten und unstrukturierten Datensätzen automatisiert erkannt und klassifiziert werden, um eine effektive Anonymisierung zu gewährleisten?

RQ2 Welche Anonymisierungsmethoden eignen sich am besten zur Wahrung der Datenqualität und zum Schutz der Privatsphäre in einem Webportal zur Datenanonymisierung?

RQ3 Wie kann kann ein Webportal entwickelt werden, das den Anonymisierungsprozess visuell darstellt, die Auswirkungen der Anonymisierung transparent erklärt und eine benutzerfreundliche Oberfläche bietet?

1.3 Abgrenzung

Diese Arbeit basiert nicht auf einer rechtlich bindenden Umsetzung im Sinne der Datenschutz-Grundverordnung der Europäischen Union (DSGVO) oder des revidiertes Datenschutzgesetz der Schweiz (revDSG). Es wird also keine rechtliche Validierung oder Kontrolle dieser Vorschriften durchgeführt.

Der Vergleich und die Evaluation von NLP-Modellen wird kein Bestandteil dieser Arbeit sein. Der Fokus liegt auf der praktischen Implementierung der automatisierten Datenanonymisierung, nicht auf der Analyse unterschiedlicher NER-Modelle. Dementsprechend wird auf vorhandene Forschung zurückgegriffen.

Die Umsetzung technischer Sicherheitsaspekte des Webportals wie die sichere Datenübermittlung von Client zu Server werden in dieser Arbeit nicht behandelt.

Für die Anonymisierung wird auf fundierte, bestehende Modelle wie k-Anonymität, ℓ-Diversität und darauf aufbauende Konzepte zurückgegriffen. Eigene Modelle werden allenfalls als Entwicklungshilfe entwickelt, werden jedoch nicht evaluiert.

Diese Arbeit schliesst zudem die visuelle Gestaltung des Webportals aus, da dieser Aspekt den Rahmen der Arbeit übersteigen würde und keine zentrale Rolle spielt. Der Fokus liegt auf der Funktionalität der automatisierten Erkennung sensibler Daten sowie der Anonymisierung.

1.4 Aufbau der Arbeit

Diese Arbeit ist in fünf Kapitel gegliedert. Die Struktur orientiert sich an Empfehlungen von Balzert, Schröder und Schäfer [5]. Dieses Buch ist ein empfohlener Leitfaden für wissenschaftliches Arbeiten an der Hochschule Luzern (HSLU).

Einleitung Das Kapitel 1 erläutert die Relevanz des Themas *automatisierte und dynamische Datenanonymisierung* und stellt die Zielstellungen dieser Arbeit vor. Ausserdem wird der Rahmen der Arbeit definiert und die Arbeitsmethodik beschrieben.

Stand der Technik Im Kapitel 2 wird auf wichtige Elemente der Erkennung sensibler Daten und der Anonymisierung eingegangen. Es werden fundierte Konzepte wie die k-Anonymität und ℓ-Diversität sowie bestehende Bibliotheken und Tools wie spaCy, Microsoft Presidio, pyCANON und Anjana vorgestellt.

Methodik Das Kapitel 3 beschreibt die methodische Umsetzung der Anforderungen dieser Arbeit. Es beinhaltet alle Entscheidungsfindungen zur Entwicklung des Webportals und der verwendeten Bibliotheken und Modellen.

Evaluation & Ergebnisse Kapitel 4 umschliesst die Bewertung des entwickelten Webportals anhand definierter Metriken. Dabei wird hauptsächlich auf die Umsetzung und Auswirkungen der Anonymisierung eingegangen.

Diskussion Im letzten Kapitel 5 werden die Ergebnisse mit den ursprünglichen Anforderungen dieser Arbeit diskutiert. Es werden die Herausforderungen, Limitationen sowie Chancen und Verbesserungspotenziale analysiert.

Für die Rechtschreibung- und Grammatikkorrektur wurde der Duden-Rechtschreibprüfer[1] als Hilfsmittel verwendet. Weiter wurde das Produkt PaperCheck[2] von HSLU-Alumni Yves Zumbühl benutzt, um ein Künstliche Intelligenz (KI)-basiertes Feedback zur Struktur der Arbeit zu erhalten.

[1] https://www.duden.de/rechtschreibpruefung-online
[2] https://papercheck.ai

1.5 Anforderungen

Die Tabelle 1.1 zeigt die aufbereiteten Anforderungen dieser Arbeit. Die Anforderungen wurden aus der Aufgabenstellung (vgl. Anhang C) und Gesprächen mit dem Betreuer Radwan Eskhita abgeleitet.

Typ	
NF	Nicht funktional
F	Funktional

Priorität	
H	Hoch
M	Mittel
N	Niedrig

Id	Anforderung	F/NF	Priorität
A1	Das Webportal soll Nutzer:innen die Möglichkeit bieten, strukturierte sowie unstrukturierte Daten hochladen zu können.	F	H
A2	Das Webportal soll sensible Daten automatisch erkennen und klassifizieren.	F	H
A3	Das Webportal soll Nutzer:innen die Möglichkeit bieten, die automatisch klassifizierten Daten manuell anpassen zu können.	F	M
A4	Das Webportal soll die k-Anonymität als Kernmethode zur Anonymisierung bereitstellen.	F	H
A5	Die Nutzer:innen sollen die Anonymisierung durch Parameter wie den k-Wert konfigurieren können.	F	M
A6	Das Webportal soll den Nutzer:innen Visualisierungen zu den Anonymisierungsauswirkungen bereitstellen.	F	H
A7	Die anonymisierten Daten sollen in verschiedenen Formaten exportiert werden können.	F	M
A8	Das Webportal soll eine intuitive und einfach bedienbare Benutzeroberfläche bieten.	NF	M
A9	Das Webportal soll effizient mit grossen Datenmengen umgehen können.	F	H
A10	Die Anonymisierung soll für eine gute Nutzer:innenerfahrung effizient durchgeführt werden.	NF	H
A11	Das Webportal soll auf verschiedenen Endgeräten und Browsern funktionieren.	NF	M
A12	Die Anonymisierungsmodelle sollen mit realitätsnahen Daten evaluiert werden.	NF	H
A13	Das Webportal soll modular erweiterbar sein, um weitere Anonymisierungsmodelle einfach integrieren zu können.	NF	M
A14	Die Erkennung sensibler Daten soll optional durch einen KI-gestützten Ansatz verfeinert werden.	F	H

Tabelle 1.1: Anforderungen

1.6 Randbedingungen

Die Arbeit wurde im Rahmen des Informatik-Studiengangs an der HSLU über einen Zeitraum von 16 Wochen und im Umfang von 12 European Credit Transfer and Accumulation System (ECTS)-Punkten durchgeführt.

1.7 Vorgehensmodell

Das Vorgehensmodell dieser Arbeit orientiert sich an der visuellen Methode Kanban. Am Anfang wurde ein Kanban-Board erstellt, welches alle Anforderungen aus Abschnitt 1.5 beinhaltet. Aus den einzelnen Anforderungen wurden Arbeitspakete definiert. Dies dient zur vereinfachten Organisation und Strukturierung der Arbeit. Weiter wurden Meilensteine definiert, um die Projektplanung zu vereinfachen und die Nachverfolgung des Fortschritts zu erleichtern.

Es gilt zu erwähnen, dass diese Arbeit eigenständig bearbeitet wurde und deshalb die Verwendung eines Vorgehensmodell sowie die Definition von Meilensteinen ausschliesslich für die persönliche Organisation verwendet wurde. Ausschnitte des Kanban-Boards sowie der Meilensteine sind im Anhang E ersichtlich.

1.8 Risiken und Herausforderungen

Für eine strukturierte Projektplanung wurde eine Risikoanalyse durchgeführt. So wurden potentielle Hindernisse frühzeitig identifiziert. Die identifizierten Risiken wurden anschliessend bewertet und für grössere Risiken wurden Minderungsmassnahmen wurden definiert. Diese Analyse basiert auf subjektiven Einschätzungen des Autors und helfen, den Projektablauf besser zu planen. Aus diesem Grund wird erst im Anhang genauer auf die Risiken eingegangen. Eine detaillierte Übersicht der Risikoanalyse ist in Anhang D zu finden.

1.9 Versionsverwaltung

Jegliche Entwicklungsartefakte werden in einem privaten GitHub-Repository verwaltet und protokolliert. Verfügbar ist das Repository unter folgendem Link[3].

[3]Wenn Sie Zugang zum Repository benötigen, wenden Sie sich bitte an den Autor.

2 Stand der Technik

Dieses Kapitel verschafft einen Überblick über die Grundlagen und bestehenden Modelle, die für die automatisierte Datenanonymisierung von Bedeutung sind. Es werden zudem relevante Metriken zur Bewertung sowie Verfahren zur Erkennung sensibler Daten vorgestellt.

2.1 Daten

In der automatisierten Datenanonymisierung spielt die Art der zu verarbeitenden Daten, insbesondere sensible Daten oder sogenannte Personally Identifiable Information (PII), eine wichtige Rolle. Grundsätzlich lassen sich die Daten in drei Kategorien einteilen:

Strukturierte Daten Strukturierte Daten sind typischerweise in tabellarischer Form gegliedert. Jede Spalte steht dabei für eine bestimmte Eigenschaft und jede Zeile bildet einen Datensatz ab. Oftmals kann eine Zeile einer Person oder einer Transaktion zugeordnet werden [6, S. 341].

Unstrukturierte Daten Als unstrukturierte Daten werden jegliche Freitexte wie E-Mails, Dokumente oder andere Nachrichten betrachtet. Sie können potentiell sensible Informationen enthalten, dies jedoch ohne definierte Struktur. Das erschwert die automatisierte Verarbeitung sowie Anonymisierung und erfordert eine kontextbasierte Verarbeitung natürlicher Sprache (NLP).

Halbstrukturierte Daten In der Praxis trifft man oft auf Mischformen der Strukturtypen. Diese Daten erfordern einen hybriden und dynamischen Ansatz, der sowohl regelbasierte als auch NLP-basierte Methoden zur Erkennung einsetzt.

2.2 Metriken

Hier werden relevante Metriken und Begriffe zur Evaluierung von NLP-Modellen sowie Anonymisierungsmodellen beschrieben.

2.2.1 Erkennung

Für diverse Metriken bezüglich der Erkennung sensibler Daten ist die folgende Wahrheitsmatrix von Relevanz [7].

Wahrheitsmatrix	
True-Positive (TP)	Das Modell hat ein positives Ergebnis korrekt vorhergesagt.
False-Positive (FP)	Das Modell hat ein negatives Ergebnis korrekt vorhergesagt.
True-Negative (TN)	Das Modell hat ein negatives Ergebnis fälschlicherweise als korrekt vorhergesagt.
False-Negative (FN)	Das Modell hat ein positives Ergebnis fälschlicherweise als inkorrekt vorhergesagt.

Tabelle 2.1: Beschreibung der Wahrheitsmatrix [7]

Präzision (engl. Precision) Die Präzision eines Modells beschreibt wie viele der positiven Ergebnisse auch korrekt klassifiziert wurden. Dabei wird die sorgfältige Selektion der wirklich korrekten Ergebnissen belohnt. False-Positive-Ergebnisse verschlechtern den Wert der Präzision [8, S. 262–263]. Die Präzision wird wie in der Gleichung 2.1 berechnet.

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

Rückruf (engl. Recall) Gegenüber der Präzision gibt der Rückruf an, wie viele aller korrekten Ergebnisse gefunden worden sind. False-Negative-Ergebnisse verschlechtern hierbei den Wert des Rückrufs [8, S. 262–263]. Der Rückruf wird wie in der Gleichung 2.2 berechnet.

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

F(1)-Mass (engl. F(1)-Score) Die Bewertungen von Präzision und Rückruf können jedoch manipuliert werden. Wenn ein korrektes Ergebnis gefunden wird und das Modell nur dieses zurückgibt, erhält man die für die Präzision die beste Bewertung. Gleichermassen geht das für den Rückruf, indem das Modell einfach alle möglichen Ergebnisse zurückgibt.

Wie Van Rijsbergen [9] beschreibt, ist es daher gängige Praxis die Präzision und den Rückruf mit einem harmonisch gewichteten Mittel zu kombinieren. Die F-Score wird wie in der Gleichung 2.3 berechnet.

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall} \tag{2.3}$$

Bei gleicher Gewichtung von Präzision und Rückruf mit $\beta = 1$ ergibt sich der bekannte F1-Score von Gleichung 2.4:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

Genauigkeit (engl. Accuracy) Anders als die Präzision bewertet die Genauigkeit den Anteil aller korrekt klassifizierten Ergebnisse. Ein perfektes Modell hätte keine falsch klassifizierte Ergebnisse und somit eine Genauigkeit von 1.0 [10]. Die Genauigkeit wird wie in der Gleichung 2.5 berechnet.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

2.2.2 Anonymisierung

Informations-Entropie Die Informationsentropie oder auch Shannon-Entropie ist eine Metrik für die Unsicherheit oder den Informationsgehalt eines Datensatzes und wurde von Shannon [11] im Jahre 1948 eingeführt. Dieses Mass beschreibt, wie viel Information nötig ist, um ein Ereignis mit einer bestimmten Wahrscheinlichkeit zu identifizieren. Die Informationsentropie wird wie in der Gleichung 2.6 berechnet.

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \tag{2.6}$$

Dabei ist X eine diskrete Zufallsvariable mit den möglichen Werten x_1, x_2, ..., x_n. Der Ausdruck $p(x_i)$ beschreibt die Wahrscheinlichkeit des Auftretens vom Ereignis x_i. Die Gleichung 2.6 verwendet die log_2, da es in Bits rechnet.

Je höher die Entropie ist, desto zufälliger und stärker verteilt sind die Ereignisse. Je niedriger die Entropie ist, desto leichter sind die Ereignisse vorherzusagen [12], [13].

Kullback-Leibler-Divergenz (KL-Divergenz) Die KL-Divergenz basiert auf der Informations-Entropie und beschreibt die Differenz zwischen zwei Wahrscheinlichkeitsverteilungen. Im Kontext dieser Arbeit misst dieses Mass wie viel Information verloren gegangen oder verzerrt wurde [14], [15]. Die KL-Divergenz wird im diskreten Fall wie in der Gleichung 2.7 berechnet.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x_i) log_2 (\frac{P(x_i)}{Q(x_i)}) \tag{2.7}$$

Je höher die KL-Divergenz ist, desto grösser ist der Unterschied der anonymisierten Daten und Originaldaten in ihrer Wahrscheinlichkeitsveteilung. Ein hoher Wert kann also zu Informationsverlust führen.

Kreuzentropie Wie Cover [13] in seinem Buch beschreibt, misst die Kreuzentropie, wie gut eine Wahrscheinlichkeitsverteilung die wahre Verteilung modelliert. Der Unterschied zur KL-Divergenz liegt darin, dass die Entropie der wahren Verteilung P auch berücksichtigt wird. Sie wird oft im Bereich des maschinellen Lernens verwendet, um zu messen, wie gut ein Modell die wahre Wahrscheinlichkeitsverteilung der Daten approximiert. Der direkte Zusammenhang zu den Gleichungen (2.6) und (2.7) ist in der Gleichung 2.8 ersichtlich.

$$D(X; P; Q;) = H(X) + D_{KL}(P||Q) \tag{2.8}$$

Das ergibt eingesetzt und vereinfacht:

$$\begin{aligned} D(X;P;Q;) &= H(X) + D_{KL}(P||Q) \\ &= -\sum_{x \in X} p(x_i) log_2 q(x_i) \end{aligned} \tag{2.9}$$

Je höher der Wert der Kreuzentropie ist, desto mehr hat sich die Datenverteilung geändert. Ein hoher Wert kann also wiederum zu Informationsverlust führen.

2.3 Erkennung sensibler Daten

Die automatisierte Erkennung sensibler Daten ist ein wesentlicher Bestandteil der späteren Datenanonymisierung. Dabei kommen verschiedene Techniken und Methoden zum Einsatz, die im Kontext dieser Arbeit grob in diese vier Kategorien unterteilt werden können:

Regulärer Ausdruck (engl. Regular Expression) (RegEx) Reguläre Ausdrücke sind beliebte Werkzeuge in der Informatik und werden auch in NLP häufig verwendet. Dabei können Zeichenmuster wie E-Mail-Adressen, Telefonnummern oder IBAN-Nummern statisch definiert werden. Diese Muster können anschliessend effizient und präzis aus beliebigen Daten identifiziert werden [16]. Friedl [17] betont in seinem Buch, dass RegEx in modernen Programmiersprachen oft mit nur wenigen Befehlen mächtige Textmanipulationen ermöglicht. Sobald kontextbasierte Wörter (z.B. identisch geschriebene Wörter mit unterschiedlichen Bedeutungen) erkennt werden sollen, funktioniert RegEx nicht mehr.

Conditional Random Fields (CRF) CRFs sind diskriminante Modelle, die basierend auf einem gegebenen Eingabewert einen Ausgabewert vorhersagen. Im Gegensatz zu generativen Modellen konzentrieren sich CRFs auf die bedingte Wahrscheinlichkeit. Wie Sutton, McCallum et al. [18] erläutern, ist dies nützlich, wenn Ausgaben voneinander abhängen. Sie erfordern jedoch manuell definierte Merkmale, was die Flexibilität einschränkt.

Named Entity Recognition NER ist eine Technik aus der Verarbeitung natürlicher Sprache (NLP), die spezifische Entitäten wie Namen oder Orte in Texten identifiziert und klassifiziert. Im Rahmen dieser Arbeit wird NER verwendet, um sensible Daten automatisch zu erkennen.

Bidirektionale kontextualisierte Sprachmodelle Bidirektional kontextualisierte Sprachmodelle sind moderne Verfahren und können semantisch komplexe Zusammenhänge verarbeiten. Sie gelten als robuster und effizienter als CRFs, da sie Kontextinformationen in beide Richtungen nutzen und keine manuell definierten Merkmale benötigen. Eine Arbeit von Naseer, Ghafoor, Khalid Alvi et al. [19] vergleicht NER-Bibliotheken und zeigt, dass Werkzeuge wie spaCy, StanfordNLP, TensorFlow und Apache OpenNLP von solchen Sprachmodellen profitieren und zunehmend in der Praxis verwendet werden.

Presidio von Microsoft ist ein praxisorientiertes Werkzeug, das auf spaCy basiert und speziell für die Erkennung und Anonymisierung von PII ausgelegt ist [20].

2.4 Techniken der Anonymisierung

Die Anonymisierung von Daten ist ein intensiv erforschter Bereich, in dem zahlreiche Techniken und Modelle entwickelt wurden.

2.4.1 Grundlagen

Zur Einordnung werden im Folgenden zentrale Begriffe des Datenschutzes knapp erläutert.

Quasi-Identifikatoren Quasi-Identifikatoren können einzeln keine Person eindeutig identifizieren. In einer Kombination mit anderen Eigenschaften können sie jedoch zur Re-Identifikation genutzt werden.

Anonymisierung Eine Anonymisierung definiert eine nicht umkehrbare Entfernung personenbezogener Attribute, sodass keine Person mehr eindeutig zugeordnet werden kann.

Pseudoanonymisierung Bei der Pseudoanonymisierung werden die sensiblen Daten durch Pseudowerte ersetzt. Oftmals wird extern eine Hilfstabelle erstellt, damit eine Re-Identifizierung der Person jederzeit vorgenommen werden kann.

2.4.2 Techniken

Im Folgenden werden Techniken zur Vorbereitung und Durchführung der Anonymisierung kurz vorgestellt.

Datenreduktion (engl. Data Reduction) Vor der Anonymisierung werden irrelevante oder nicht notwendige Attribute ganz entfernt, um die Datenmenge zu verkleinern und das Re-Identifikationsrisiko zu senken. Dies steht im Einklang mit dem Prinzip der Datenminimierung (vgl. Art. 5 Abs. 1, DSGVO).

Maskierung (engl. Masking) Ersetzt Teile sensibler Informationen durch Platzhalter, sodass der Originalwert unkenntlich gemacht wird. Dies wird häufig in Logs verwendet.

Unterdrückung (engl. Suppression) Anders als bei der Maskierung werden bei der Unterdrückung gezielt Werte vollständig entfernt, wenn diese ein zu hohes Risiko für eine Re-Identifikation darstellen. Diese Technik wird oft kombiniert mit der Generalisierung.

Generalisierung (engl. Generalization) Reduziert den Detailgrad von Attributen, beispielsweise durch Altersgruppen oder Postleitzahlenbereiche. Dadurch entstehen Äquivalenzklassen zur Erreichung von k-Anonymität.

Verfälschung (engl. Perturbation) Verändert Werte gezielt (z.B. durch zufälliges Rauschen), um Originaldaten zu verschleiern, ohne den statistischen Gesamteindruck zu verlieren. Dies ist die Grundlage der differenziellen Privatsphäre.

2.4.3 Modelle

Im Verlauf der Jahre haben sich im Bereich der Datenanonymisierung verschiedene Modelle etabliert. Diese bieten unterschiedliche Schutzstufen gegen eine Re-Identifikation. Shamsinejad, Banirostam, Pedram et al. [21, S. 257] sowie Sáinz-Pardo Díaz und López García [22] fassen bewährte Methoden zusammen, die auch in diesem Projekt berücksichtigt werden.

k-Anonymität (engl. k-Anonymity) Die im Jahre 1998 entwickelte k-Anonymität [23] stellt sicher, dass jedes Objekt innerhalb eines Datensatzes durch mindestens k anderen unterscheidbar ist. Dieses Modell gilt als grundlegendes Konzept der Anonymisierung, auf dem viele Erweiterungen aufbauen [24].

(α, k)-Anonymität (engl. (α, k)-Anonymity) Dies ist eine Erweiterung der k-Anonymität, bei der zusätzlich ein Mass für Diversität innerhalb der Gruppen integriert wird. Dabei wird versucht, dominante Werte in Äquivalenzklassen zu reduzieren. Der Parameter α legt den maximal zulässigen Anteil eines sensiblen Werts fest [25].

ℓ-Diversität (engl. ℓ-Diversity) Machanavajjhala, Kifer, Gehrke et al. [26] entwickelten mit der ℓ-Diversität ein Modell, das Gruppen mit je ℓ verschiedenen, semantisch relevanten Werten für ein sensibles Attribut erstellt. So wird sichergestellt, dass keine homogenen Gruppen bestehen.

Entropie-ℓ-Diversität (engl. Entropy-ℓ-Diversity) Als Verfeinerung der ℓ-Diversität wird die Informations-Entropie integriert, um die Wahrscheinlichkeitsverteilung der sensiblen Daten innerhalb Äquivalenzklassen zusätzlich zu bewerten [26].

Rekursive-(c, ℓ)-Diversität (engl. Recursive-(c, ℓ)-Diversity) Die rekursive (c, ℓ)-Diversität stellt sicher, dass innerhalb einer Äquivalenzklasse nicht nur mehrere verschiedene sensible Werte vorliegen, sondern dass kein einzelner Wert zu häufig vorkommt. Dazu wird überprüft, ob der häufigste Wert nicht deutlich häufiger auftritt als die folgenden $(\ell - 1)$ häufigsten Werte, gemessen durch ein Verhältnis c. Dies schützt gegen Attribute mit dominanten sensiblen Ausprägungen [15].

Einfache-β-Ähnlichkeit (engl. Basic-β-Likeness) Die einfache β-Ähnlichkeit misst, wie stark sich die Verteilung sensitiver Werte in einer Äquivalenzklasse von der Gesamtverteilung unterscheidet. Sie erlaubt eine Bewertung des Informationslecks auf Basis einer logarithmischen Schranke [27].

Erweiterte-β-Ähnlichkeit (engl. Enhanced-β-Likeness) Die erweiterte β-Ähnlichkeit verschärft die Bedingungen des vorherigen Modells, indem sie auch den Schutz häufig auftretender sensibler Werte berücksichtigt. Dies ist besonders relevant, wenn bestimmte Werte in der Grundverteilung überrepräsentiert sind und daher gezielt geschützt werden müssen.

T-Nähe (engl. T-Closeness) Das T-Nähe-Modell von Li, Li und Venkatasubramanian [15] verlangt, dass die Verteilung sensitiver Daten innerhalb einer Gruppe nicht stark von der Gesamtverteilung abweicht. So können Rückschlüsse aufgrund untypischer Verteilungen verhindert werden.

Differenzielle Privatsphäre (engl. Differential Privacy) Die differenzielle Privatsphäre verfolgt einen mathematischen Ansatz, bei dem gezielt Rauschen zu den Daten hinzufügt wird. Dabei wird das Ziel verfolgt, möglichst genaue statistische Auswertungen zu ermöglichen, ohne die Identität offenzulegen [28]. Die Methode von Dwork, Roth et al. [29] ist heute ein zentraler Bestandteil modernen Datenschutzmethoden.

Hinweis: Einige der beschriebenen Modelle (z.B. *k-Anonymität* oder *ℓ-Diversität*) tragen die gleichen Namen wie die entsprechenden Bewertungsmetriken, die von der Bibliothek pyCANON zur Verfügung gestellt werden. Auf diese Metriken wird in Abschnitt 4.7 genauer eingegangen.

3 Methodik

Im folgenden Kapitel wird das methodische Vorgehen zur Entwicklung des Webportals zur automatisierten Datenanonymisierung erläutert. Es wird auf die ausgewählten Modelle und Technologien eingegangen, die im Prototyp verwendet werden. Die Methodik orientiert sich an den formulierten Forschungsfragen und bildet das Fundament für die spätere Umsetzung und Diskussion.

3.1 Architektur des Webportals

Die Architektur des Webportals besteht aus zwei Komponenten. Im Backend werden die Logiken zur Erkennung sensiblen Daten sowie für die unterschiedlichen Anonymisierungsmodelle implementiert. Im Frontend werden die Interaktionen der Nutzer:innen abgehandelt und Visualisierungen präsentiert. In dieser Sektion werden die Entscheidungsfindungen für Back- und Frontend beschrieben.

Die Auswahl der Technologien erfolgt durch einen Vergleich mehrerer Kriterien, die aus den in Tabelle 1.1 definierten funktionalen und nicht-funktionalen Anforderungen abgeleitet wurden. Solche manuellen Vergleiche sind oft komplex, insbesondere in einer begrenzten Teamgrösse. Im Rahmen dieser Bachelorarbeit wird dieser Prozess eigenständig durchgeführt. Zur systematischen und transparenten Bewertung wird die Nutzwertanalyse von Six Sigma Black Belt [30] angewendet. Sie ermöglicht eine objektive Gewichtung und Bewertung der Kriterien, was eine fundierte Entscheidungsfindung unterstützt. Ein paarweiser Vergleich ist hierbei besonders hilfreich, da er die Bewertung der Kriterien vereinfacht und den Entscheidungsprozess klarer strukturiert [31].

Obwohl diese wissenschaftlich fundierten Methoden eine objektive Grundlage bieten, bleibt eine gewisse Subjektivität in der Gewichtung der Kriterien bestehen. Da viele Faktoren nicht immer eindeutig messbar sind, kann dies die Entscheidungsfindung beeinflussen.

3.1.1 Entscheidungsfindung Programmiersprache

Die Entscheidungsfindung der Programmiersprache ist eine zentrale Grundlage für die technische Umsetzung der automatisierten Anonymisierung. Um eine fundierte Entscheidung treffen zu können, wurden folgende Kriterien und deren Relevanz definiert:

K1 - Eignung für Datenanonymisierung Die Programmiersprache sollte eine breite Unterstützung für bestehende Anonymisierungsmodelle bieten. Dazu zählen Bibliotheken, Frameworks und Schnittstellen zur effizienten Umsetzung anerkannter Modelle wie k-Anonymität oder ℓ-Diversität.

K2 - Flexibilität für KI-Integration Da die Integration von KI-gestützten Ansätzen zur Verbesserung der Erkennung und Anonymisierung denkbar ist, sollte die Programmiersprache eine mühelose Anbindung ermöglichen.

K3 - Leistungsfähigkeit bei grossen Datenmengen Das Webportal soll auch bei stetig wachsenden Datenmengen in der Lage sein, diese effizient zu verarbeiten. Die Sprache

sollte so gewählt werden, dass die Skalierbarkeit kein Hindernis darstellt.

K4 - Erweiterbarkeit der Architektur Die Programmiersprache soll die Verwendung flexibler und erweiterbarer Softwarearchitekturen ermöglichen. Dazu gehören beispielsweise die Einbindung externer Schnittstellen oder die Nutzung von Entwurfsmustern.

K5 - Unterstützung für wissenschaftliche Visualisierung Das Webportal soll die Auswirkungen der Anonymisierung visualisieren können. Deshalb ist eine Sprache mit integrierten Visualisierungsmöglichkeiten vorteilhaft. Häufig handelt es sich dabei um Programmiersprachen, die sowohl im Frontend als auch im Backend eingesetzt werden können.

K6 - Modularität und Wartbarkeit Eine klare Trennung durch Modularität führt zu einer losen Kopplung der Komponenten des Webportals. Dies ist entscheidend für die langfristige Wartbarkeit und Erweiterbarkeit.

K7 - Community und verfügbare Bibliotheken Eine aktive Community erleichtert die Entwicklung und die Problembehebung. Mit einer grossen Community geht oft eine umfassende und gut strukturierte Dokumentation der Programmiersprache einher. Dies kann für die Entwicklung sehr nützlich sein.

K8 - Entwicklungs- und Debugging-Freundlichkeit Die Entwicklung des Webportals sollte nicht durch eine zu hohe Komplexität der Sprache erschwert werden. Sie sollte eine verständliche Syntax und gute Debugging-Werkzeuge bieten.

K9 - Sicherheit Obwohl die Sicherheit nicht im Fokus dieser Arbeit steht, sollte die Programmiersprache keine bekannten Sicherheitsmängel aufweisen.

K10 - Kompatibilität mit Webtechnologien Die Ergebnisse dieser Arbeit sollen in einem Webportal dargestellt werden. Aus diesem Grund ist die Kompatibilität mit Webtechnologien essenziell.

Die Ergebnisse dieser Entscheidungsfindung werden in Abschnitt 4.1 thematisiert.

3.1.2 Entscheidungsfindung Frontend

Folglich werden die Kriterien für die Entscheidungsfindung der Frontend-Technologie und deren Relevanz erläutert:

K1 - Einfachheit der Entwicklung Das Frontend sollte eine einfache und schnelle Entwicklung ermöglichen, damit der Fokus auf den Funktionalitäten der automatisierten Anonymisierung liegt.

K2 - Unterstützung für interaktive Visualisierungen Die Ergebnisse der Anonymisierung sollen visuell dargestellt werden. Die verwendete Technologie soll daher interaktive Visualisierungen wie Diagramme und Tabellen ermöglichen.

K3 - Performance für Datenverarbeitung Das Frontend sollte ebenso wie die Programmiersprache gut geeignet sein, grosse Datenmengen effizient zu verarbeiten.

K4 - Flexibilität und Erweiterbarkeit Die Frontend-Technologie sollte für zukünftige Entwicklungen gut erweiterbar und flexibel einsetzbar sein.

K5 - User-Interaktion Intuitive Interaktionen mit den Nutzer:innen sollen von der Frontend-Technologie unterstützt werden. Interaktionen wie das Hochladen von Daten oder die Navigation durch die Erkennung und Anonymisierung sind wichtige Bestandteile des Webportals.

K6 - Unterstützung für Datenexport Die Technologie soll das Herunterladen von Dateien in den gewünschten Formaten unterstützen.

K7 - Ressourcenverbrauch Das Frontend sollte ressourcenschonend sein und keine hohe Rechenlast auf dem Client verursachen.

K8 - Unterstützung für Prototyping Eine Unterstützung für schnelles Prototyping ist von Vorteil, da das resultierende Webportal keine produktive Umgebung darstellt.

K9 - Frontend-Anpassbarkeit Obwohl der Fokus auf den funktionalen Aspekten der Anonymisierung liegt, sollte das Frontend dennoch anpassbar sein, um eine bessere Nutzer:innenführung zu ermöglichen.

K10 - Lernkurve Im kleinen Rahmen dieser Bachelorarbeit erleichtert eine niedrige Einstiegshürde die Implementierung. Es sollten keine umfangreichen Frontendkenntnisse vorausgesetzt werden.

Die Ergebnisse dieser Entscheidungsfindung werden in Abschnitt 4.2 thematisiert.

3.2 Evaluation der Anonymisierungsmodelle

Für die Evaluation der implementierten Anonymisierungsmodelle wird die Python-Bibliothek pyCANON verwendet. pyCANON wurde von Sáinz-Pardo Díaz und López García [22] entwickelt und bietet eine umfassende Sammlung etablierter Metriken zur Bewertung von Datenschutz und Anonymität. Dazu zählen unter anderem k-Anonymität, ℓ-Diversität und t-Closeness.

Die Nutzung von pyCANON ermöglicht eine fundierte, objektive und reproduzierbare Bewertung der Anonymisierungsergebnisse. Durch die Verwendung eines wissenschaftlich validierten Werkzeugs wird sichergestellt, dass die Qualität der Anonymisierung anhand standardisierter Kennzahlen messbar ist und somit vergleichbare Aussagen über die Wirksamkeit der verschiedenen Modelle getroffen werden können.

Im Kontext dieser Bachelorarbeit dienen die Bewertungen von pyCANON dazu, die Stärken und Schwächen der eingesetzten Anonymisierungsverfahren aufzuzeigen.

3.3 Umsetzung der Anonymisierungsmodelle

Für die technische Umsetzung der Anonymisierungsmodelle wurde die Python-Bibliothek Anjana verwendet. Diese wurde von Sáinz-Pardo Díaz und López García [32] entwickelt und baut auf den in pyCANON definierten Metriken auf. Anjana ermöglicht somit eine direkte, praktische Anwendung der Anonymisierungsmodelle auf Basis wissenschaftlich validierter Kriterien.

Die Verwendung von Anjana erlaubt eine risikoarme und nachvollziehbare Umsetzung der im Abschnitt 3.2 beschriebenen Bewertungsmethoden. Die modulare Architektur des entwickelten Systems unterstützt dabei nicht nur die Einbindung bestehender Verfahren, sondern erlaubt bei Bedarf auch die Erweiterung um eigene Anonymisierungsmodelle.

Da Anjana auf pyCANON basiert, ergibt sich eine konsistente Kombination von Implementierung und Evaluation. Dadurch kann die Qualität der Anonymisierung direkt im Anschluss an den Anonymisierungsprozess objektiv gemessen werden. Diese Evaluationen werden im 5 weiter diskutiert.

Die nachfolgende Abbildung 3.1 veranschaulicht die Beziehung zwischen Anjana und pyCANON.

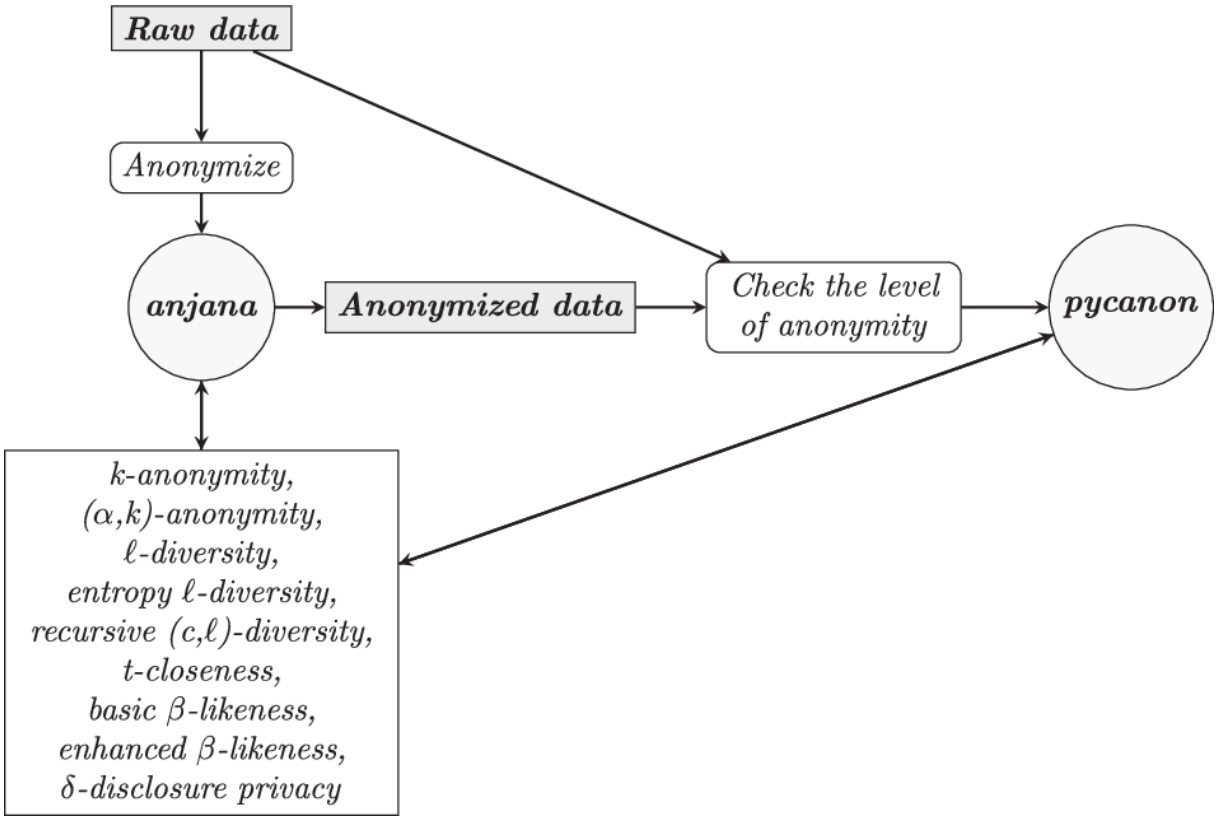


Abbildung 3.1: Übersicht Anjana / pyCANON

3.4 Testkonzept

Für dieses Projekt ist ein frühzeitig definiertes und durchdachtes Testkonzept zentral, um die Qualität der Erkennung und Anonymisierung laufend zu überwachen. Deshalb wurden geeignete Tests definiert und implementiert, um Optimierungen oder Verschlechterungen laufend zu erkennen.

Die folgenden Testarten wurde definiert:

Bewertung der Erkennung des dreistufigen Prozesses Die automatische Erkennung sensibler Daten wird durch ein vordefiniertes Beispiel überprüft. Die erwarteten Identifikatoren werden festgelegt und sollen anschliessend von der Erkennungskomponente korrekt erkannt werden.

Bewertung der Anonymisierung mit pyCANON Die verschiedenen Anonymisierungsmodelle werden mithilfe von pyCANON bewertet und miteinander verglichen. Dabei liegt ein besonderer Fokus auf dem systematischen Vergleich der Informationsentropie der Attribute.

Performanz-Tests Mit gezielten Performanz-Tests können einzelne Anonymisierungsmodelle bewertet werden. Zudem wird die Verarbeitungszeit von unterschiedlich grossen Testdatensätzen im Hinblick auf die Skalierbarkeit evaluiert.

3.5 Visualisierung der Anonymisierungseffekte

Um die Auswirkungen der angewandten Anonymisierung transparent und nachvollziehbar darzustellen, werden im Webportal verschiedene Visualisierungstypen eingesetzt. Ziel ist es, den Grad der Veränderung sowie den damit verbundenen Datenschutzgewinn verständlich zu präsentieren. Die gewählten Visualisierungen sollen nicht nur die Wirksamkeit der Anonymisierung zeigen, sondern auch Rückschlüsse auf die verbleibende Datenqualität offenlegen.

Die Umsetzung erfolgt mit Python-Bibliotheken wie *matplotlib* und *seaborn* für die grafische Darstellung sowie mit dem Tool pyCANON zur Berechnung der Metriken. Drei Hauptansätze werden verfolgt:

- **Direkter Vorher-Nachher-Vergleich:** Die Original- und anonymisierten Daten werden gegenübergestellt, um die Veränderungen durch die Anonymisierung sichtbar zu machen. Diese Visualisierungen zeigen, welche Attribute generalisiert oder unterdrückt wurden.
- **Entropie-Analyse:** Durch die Informations-Entropie aus Unterabschnitt 2.2.2 wird die Datenvielfalt vor und nach der Anonymisierung aufgezeigt. Dies ist ein Hinweis auf den Grad der Anonymisierung und den damit verbundenen Informationsverlust.
- **Metrikbasierte Bewertung mit pyCANON:** Zur metrikbasierten Bewertung werden die Vordefinierten Konzepte von pyCANON verwendet. Diese Konzepte sowie die Ergebnisse dazu werden in Unterabschnitt 2.4.3 und in Abschnitt 4.7 genauer beschrieben.

3.6 Erkennung sensibler Daten

Für die Erkennung der potentiell sensiblen Daten wurden spaCy und Microsoft Presidio eingesetzt. Diese Produkte sind weit verbreitet sowie gut erforscht und können für strukturierte- und unstrukturierte Daten eingesetzt werden. Sie ermöglichen eine automatisierte Identifikation und Klassifikation durch NER. In der Abbildung 3.2 wird der Erkennungsprozess von Microsoft Presidio gezeigt. Da dieser Bereich in dieser Arbeit jedoch nicht Gegenstand der Bewertung oder Analyse ist, erfolgt keine detaillierte Betrachtung dieser Komponente.

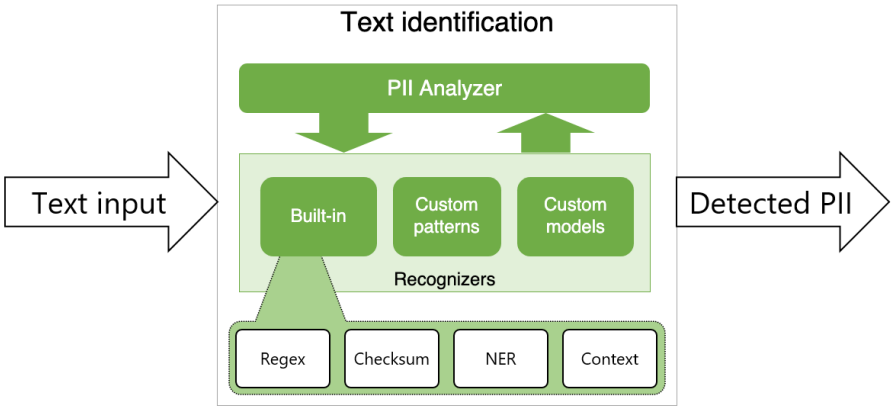


Abbildung 3.2: Microsoft Presidio [33]

3.7 Softwarearchitektur und Entwicklungsprinzipien

Das Webportal wurde modular aufgebaut, damit die Kernfunktionen klar voneinander getrennt sind. So sind beispielsweise die Komponenten zur Erkennung sensibler Daten, zur Anwendung der Anonymisierungsmodelle sowie zur Visualisierung der Ergebnisse voneinander entkoppelt implementiert. Durch diese Trennung wird eine einfache Wartbarkeit und zukünftige Erweiterung ermöglicht.

Ein weiteres Augenmerk wurde auf die saubere Kommentierung des Codes gelegt. Alle entwickelten Komponenten und Funktionen wurden mit erklärenden Kommentaren versehen, um die Nachvollziehbarkeit für Entwickler:innen zu erhöhen. Dies unterstützt zukünftige Weiterentwicklungen und eine effiziente Fehlersuche.

Die Anonymisierungsmodelle wurden bewusst modular nach dem Prinzip der *Separation of Concerns* [34] implementiert. Diese Struktur erlaubt es, neue Modelle wie die differenzielle Privatsphäre problemlos in das bestehende System zu integrieren. Auch erlaubt sie die einfache Erweiterung um weitere Anonymisierungsbiliotheken oder den Austausch bestehender Lösungen wie Anjana und pyCANON.

3.8 Datenfluss

Die Abbildung 3.3 zeigt den modularen Datenfluss der automatisierten Datenanonymisierung.

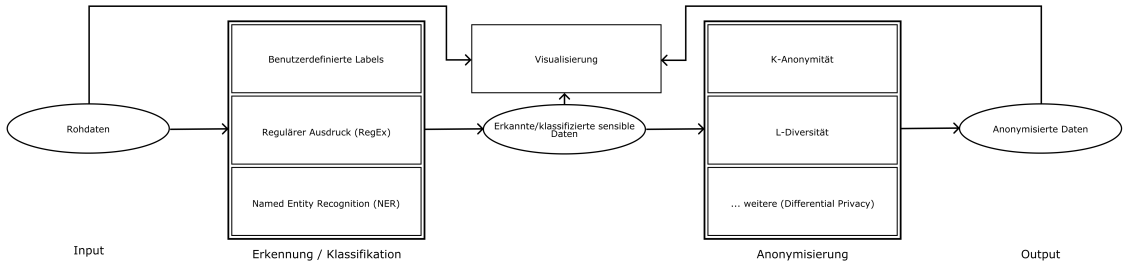


Abbildung 3.3: Datenfluss

Der Ablauf ist in vier Schritte gegliedert:

1. **Dateneingabe:** Die Nutzer:innen laden strukturierte, unstrukturierte oder halbstrukturierte Daten (z.B. Comma-Separated Values (CSV), JavaScript Object Notation (JSON), Extensible Markup Language (XML) oder TXT) in das System.

2. **Erkennung / Klassifikation:**
   Dieser Schritt umfasst drei Mechanismen zur Identifikation potenziell sensibler Daten und zur Bestimmung des *sensitivity score*:

   - **Benutzerdefinierte Labels:** Durch benutzerdefinierte Labels können domänenspezifische Schlagwörter definiert werden, die in der Erkennung fokussiert werden sollen.
   - **RegEx:** Reguläre Ausdrücke untersuchen die Daten effizient nach bekannten Mustern (z.B. International Bank Account Number (IBAN)'s oder Telefonnummern)
   - **NER:** Mithilfe von NLP-Methoden wie spaCy oder Microsoft Presidio werden Entitäten inklusive der Kontexte erkannt.

3. **Anonymisierung:**
   Die erkannten sensiblen Daten werden mit einem gewünschten Modell anonymisiert. Die Auswahl erfolgt entweder manuell oder automatisch (vgl. Abschnitt 5.7).

4. **Visualisierung:**
   Der gesamte Anonymisierungsprozess wird durch eine Visualisierung begleitet. Die Nutzer:innen erhalten transparente Einsicht in die Rohdaten, die Erkennung der sensiblen Daten, die anonymisierten Daten sowie diverse Metriken zur Anonymisierung.

3.9 Anwendungsfallanalyse

Aus den Anforderungen von Tabelle 1.1 ergeben sich folgende Anwendungsfälle:

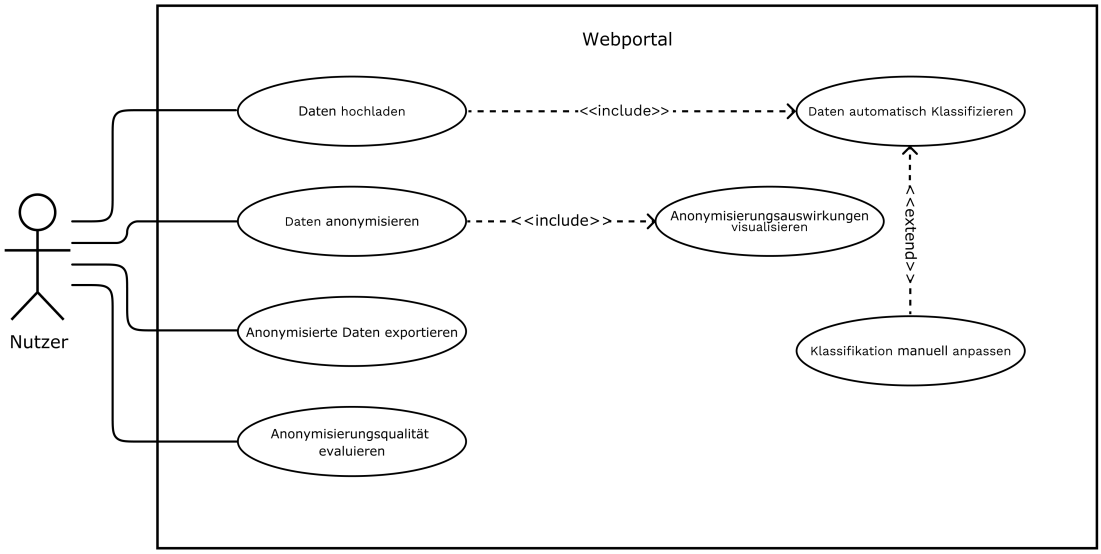


Abbildung 3.4: Anwendungsfälle

UC1 - Daten hochladen (A1-A3, A9, A11, A14) Der/die Nutzer:in kann Daten hochladen und diese werden automatisch identifiziert und klassifiziert. Bei Bedarf kann der/die Nutzer:in die Klassifikation manuell anpassen.

UC2 - Daten anonymisieren (A4-A6, A10, A11, A13) Der/die Nutzer:in kann die hochgeladenen Daten mit dem vorgeschlagenen oder dem ausgewählten Modell anonymisieren. Die Auswirkungen der Anonymisierung werden dem/der Nutzer:in visuell präsentiert.

UC3 - Anonymisierte Daten exportieren (A7, A11) Der/die Nutzer:in kann die anonymisierten Daten als CSV-, XML- oder JSON-Datei exportieren.

UC4 - Anonymisierungsqualität evaluieren (A6, A8, A11, A12) Der/die Nutzer:in kann über diverse Grafiken die Qualität der Anonymisierung evaluieren.

3.10 Methodische Zusammenfassung

Im Rahmen dieser Bachelorarbeit wurde ein systematisches Vorgehen ausgewählt, um eine fundierte und nachvollziehbare Umsetzung der automatisierten und dynamischen Datenanonymisierung zu realisieren. Die Entscheidungsfindung der Technologien erfolgte mittels einer Nutzwertanalyse unter Berücksichtigung der Anforderungen aus Abschnitt 1.5.

Die Umsetzung basiert auf einer modularen Architektur. So wird eine flexible Erweiterung durch weitere Komponenten erlaubt. Bewährte Anonymisierungsmethoden werden auf der Grundlage von Anjana in das Projekt integriert. Evaluationen der Qualität der Anonymisierung werden mit der darunterliegenden Bibliothek pyCANON vorgenommen.

Zur Qualitätssicherung und Evaluierung wurde ein umfassendes Testkonzept definiert, das Unit-Tests und Performanzmessungen bei unterschiedlichen Datensatzgrössen einschliesst. Mit verschiedenen Visualisierungstypen wird versucht, die Effekte der Anonymisierung nachvollziehbar zu präsentieren.

4 Evaluation & Ergebnisse

In diesem Kapitel werden die Ergebnisse sowie die dazugehörigen Evaluationen erläutert. Grundlage dafür ist das in Kapitel 3 beschriebene methodische Vorgehen. Es werden sowohl die Entscheidungsfindungen als auch die Bewertungen der Anonymisierungsmodelle sowie ergänzende Messungen und Visualisierungen dokumentiert.

4.1 Evaluation der Programmiersprache

Die in Unterabschnitt 3.1.1 beschriebenen Kriterien zur Entscheidungsfindung der Programmiersprache wurden mithilfe eines paarweisen Vergleichs gegenübergestellt. Dabei wird das Gewicht der einzelnen Kriterien für die Berechnung einer prozentualen Wichtigkeit verwendet:

Wichtiger \ **als**	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	**Summe**	**%**
K1	•	1	1	1	1	1	1	1	1	1	9	20.00%
K2	0	•	0	0	1	1	1	1	1	0	5	11.11%
K3	0	1	•	1	1	1	1	1	1	1	8	17.78%
K4	0	1	0	•	1	1	0	0	1	0	4	8.89%
K5	0	0	0	0	•	0	0	0	1	0	1	2.22%
K6	0	0	0	0	1	•	1	0	1	0	3	6.67%
K7	0	0	0	1	1	0	•	0	1	0	3	6.67%
K8	0	0	0	1	1	1	1	•	1	0	5	11.11%
K9	0	0	0	0	0	0	0	0	•	0	0	0.00%
K10	0	1	0	1	1	1	1	1	1	•	7	15.56%

Tabelle 4.1: Entscheidungsfindung Programmiersprache - Paarweiser Vergleich [30]

Für die Nutzwertanalyse wurden weit verbreitete Programmiersprachen evaluiert. Die Bewertung basiert auf Berichten, die Trends und die Nutzung von Programmiersprachen in Versionsverwaltungen vom Jahr 2024 analysieren [35], [36], [37].

Die vier führenden Programmiersprachen wurden anschliessend in Tabelle 4.2 mit einer Bewertungszahl $B(0-10)$ bewertet. Aus dem Produkt der Gewichtung und der Bewertungszahl des Kriteriums resultiert der Wert W. Das Kriterium 9 hat in dieser Nutzwertanalyse eine Gewichtung von 0 und wird deshalb nicht in die Entscheidung einfliessen, da die interne Sicherheit des Webportals in dieser Arbeit keine Relevanz hat.

	TypeScript (JavaScript)		**Java**		**C#/.NET**		**Python**	
	B	*W*	*B*	*W*	*B*	*W*	*B*	*W*
K1 (20.00%)	5	1.00	7	1.40	6	1.20	9	1.80
K2 (11.11%)	4	0.44	6	0.67	5	0.56	9	1.00
K3 (17.78%)	6	1.07	8	1.42	7	1.24	6	1.07
K4 (8.89%)	7	0.62	8	0.71	8	0.71	8	0.71
K5 (2.22%)	3	0.07	5	0.11	4	0.09	10	0.22
K6 (6.67%)	8	0.53	7	0.47	7	0.47	8	0.53
K7 (6.67%)	8	0.53	9	0.60	7	0.47	10	0.67
K8 (11.11%)	8	0.89	6	0.67	7	0.78	8	0.89
K9 (0.00%)	6	-	9	-	9	-	7	-
K10 (15.56%)	10	1.56	7	1.09	8	1.24	7	1.09
Summe	**-**	6.71	**-**	7.13	**-**	6.76	**-**	7.98

Tabelle 4.2: Entscheidungsfindung Programmiersprache - Nutzwertanalyse [30]

4.1.1 Entscheidung

Insbesondere durch die Kriterien K1, K2, K5 und K7 sticht Python in der Tabelle 4.2 als die geeignetste Programmiersprache heraus und wird somit für diese Arbeit verwendet.

4.2 Evaluation der Frontend-Technologien

Wie in Abschnitt 4.1 wird der gleiche Prozess nochmals für die Entscheidungsfindung der Frontend-Technologie durchgeführt. Dabei wurden Kriterien (vgl. Unterabschnitt 3.1.2) definiert und mithilfe eines paarweisen Vergleichs gegenübergestellt. In Tabelle 4.3 wird das Gewicht der einzelnen Kriterien für die Bestimmung einer prozentualen Wichtigkeit verwendet:

Wichtiger \ **als**	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	**Summe**	**%**
K1	•	0	0	0	0	0	0	0	0	1	1	2.22%
K2	1	•	0	1	0	0	1	0	1	1	5	11.11%
K3	1	1	•	1	1	1	1	1	1	1	9	20.00%
K4	1	0	0	•	0	0	0	0	1	0	2	4.44%
K5	1	1	0	1	•	0	0	0	1	1	5	11.11%
K6	1	1	0	1	1	•	0	0	1	1	6	13.33%
K7	1	0	0	1	1	1	•	1	1	1	7	15.56%
K8	1	1	0	1	1	1	0	•	1	1	7	15.56%
K9	1	0	0	0	0	0	0	0	•	0	1	2.22%
K10	0	0	0	1	0	0	0	0	1	•	2	4.44%

Tabelle 4.3: Entscheidungsfindung Frontend - Paarweiser Vergleich [30]

Basierend auf der Entscheidung zur Programmiersprache wurden für die Nutzwertanalyse vier etablierte Python-Frontend-Technologien evaluiert: Streamlit, Django, FastAPI und Streamlit. In der Arbeit von Akkem, Kumar und Varanasi [38] wurde ein ausführlicher Vergleich zwischen Streamlit, Flask und Django durchgeführt. Streamlit stach dabei heraus und wirkt nach Einarbeitung in eine detaillierte Lektüre passend für diese Arbeit [39]. Dennoch wird für eine fundierte Entscheidungsfindung zusätzlich eine Nutzwertanalyse durchgeführt. Die Technologie FastAPI wurde zusätzlich in diese Analyse aufgenommen, da sie immer mehr an Bedeutung gewinnt.

Die vier Technologien wurden darauffolgend mit einer Bewertungszahl $B(0-10)$ bewertet. Aus dem Produkt der Gewichtung und der Bewertungszahl des Kriteriums resultiert der Wert W. In der Tabelle 4.4 zeigt sich Streamlit als die geeignetste Frontend-Technologie.

	Streamlit		**Flask**		**FastAPI**		**Django**	
	B	*W*	*B*	*W*	*B*	*W*	*B*	*W*
K1 (2.22%)	10	0.22	6	0.13	7	0.16	5	0.11
K2 (11.11%)	10	1.11	5	0.56	6	0.67	7	0.78
K3 (20.00%)	7	1.40	9	1.80	9	1.80	8	1.60
K4 (4.44%)	4	0.18	9	0.40	9	0.40	8	0.36
K5 (11.11%)	9	1.00	7	0.78	8	0.89	7	0.78
K6 (13.33%)	9	1.20	8	1.07	8	1.07	7	0.93
K7 (15.56%)	9	1.40	9	1.40	8	1.24	6	0.93
K8 (15.56%)	10	1.56	7	1.09	7	1.09	5	0.78
K9 (2.22%)	4	0.09	7	0.16	8	0.18	9	0.20
K10 (4.44%)	10	0.44	7	0.31	7	0.31	6	0.27
Summe	-	8.60	-	7.69	-	7.80	-	6.73

Tabelle 4.4: Entscheidungsfindung Frontend - Nutzwertanalyse [30]

4.2.1 Entscheidung

Durch die Resultate der Arbeit von Akkem, Kumar und Varanasi [38] sowie der Nutzwertanalyse aus Tabelle 4.4 wird das Webportal mit der Frontend-Technologie Streamlit implementiert.

Diese Entscheidung wird auch durch aktuelle Forschung gestützt. Akkem, Kumar und Varanasi [38, S. 4695] stellen Streamlit als besonders geeignet für Datenprojekte und schnelle Frontendentwicklung dar, insbesondere durch die native Unterstützung von Python. Ebenso heben Khorasani, Abdou und Fernández [39] hervor, dass Streamlit im Vergleich zu traditionellen Technologien wie Flask oder React besonders prototyping-freundlich ist und weniger technisches Vorwissen im Frontend-Bereich erfordert.

Hinweis: Eine detaillierte Installations- und Bedienungsanleitung des Prototyps befindet sich im Anhang A und Anhang B.

4.3 Realitätsnaher Datensatz

Für die Anwendung und Evaluation der integrierten Anonymisierungsmodellen wurde ein realitätsnaher Beispieldatensatz verwendet. Dabei handelt es sich um den Datensatz `example.csv` aus der Beispielsammlung von *ARX*[1], die typischerweise für Datenschutz- und Anonymisierungsstudien verwendet wird.

[1] https://arx.deidentifier.org

Der Datensatz umfasst 11 Attribute und bildet typische personenbezogene Informationen ab. Die Attribute lauten wie folgt:

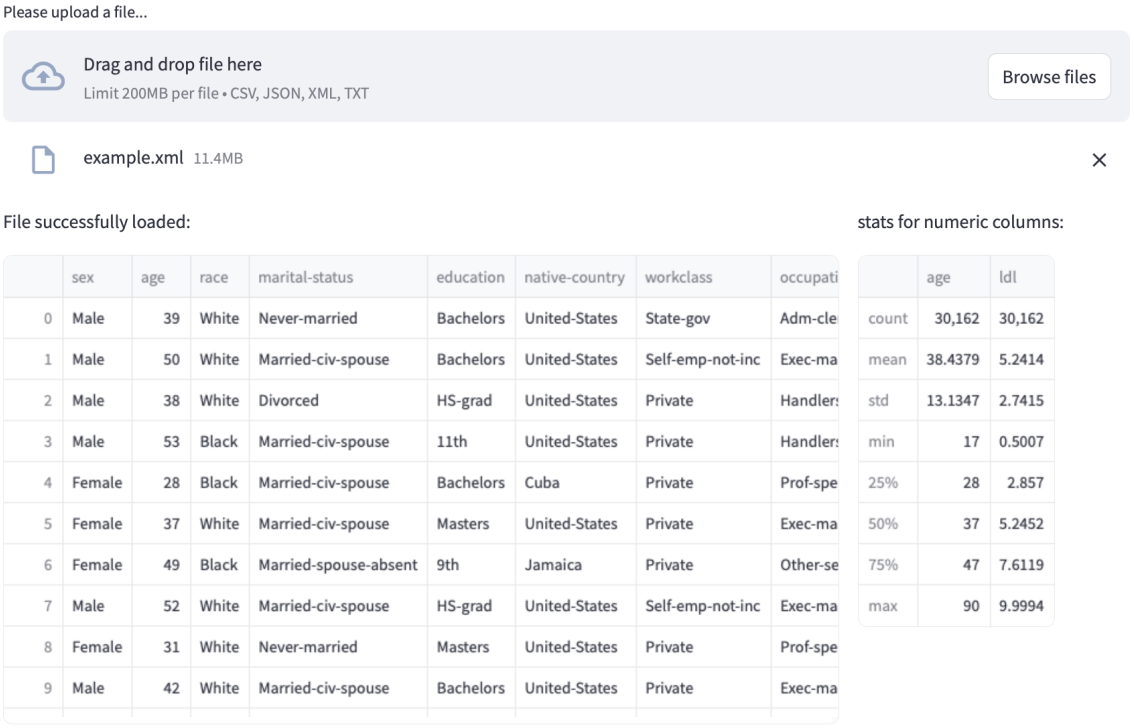
- `sex`
- `age`
- `race`
- `marital-status`
- `education`
- `native-country`
- `workclass`
- `occupation`
- `salary-class`
- `ldl` (Low-Density Lipoprotein, ein medizinischer Wert)

Die enthaltenen Daten bieten ein realistisches Szenario zur Anonymisierung sensibler personenbezogener Informationen. Sie beinhalten sowohl numerische als auch kategoriale Werte und eignen sich damit gut zur Evaluation der integrierten Anonymisierungsmodelle wie *k-Anonymität*, *ℓ-Diversität* und *t-Closeness*.

Die Struktur und Realitätsnähe des Datensatzes machen ihn zu einem idealen Ausgangspunkt für die praxisnahe Demonstration und Bewertung der Erkennung sensibler Daten sowie Anonymisierung innerhalb des Webportals.

4.4 Daten-Upload

Im ersten Schritt der Datenanonymisierung auf dem Webportal kann ein beliebiger Datensatz hochgeladen werden. Unterstützt werden strukturierte, halbstrukturierte sowie unstrukturierte Formate, darunter CSV, JSON, XML und reine Textdateien (TXT). Nach dem Upload erscheint eine Vorschau der Rohdaten. Für strukturierten Daten erscheinen zusätzlich statistische Kennzahlen zu numerischen Attributen. Die Abbildung 4.1 zeigt die Vorschau des in Abschnitt 4.3 beschriebenen Datensatzes inklusive berechneter Statistik.



	sex	age	race	marital-status	education	native-country	workclass	occupati
0	Male	39	White	Never-married	Bachelors	United-States	State-gov	Adm-cle
1	Male	50	White	Married-civ-spouse	Bachelors	United-States	Self-emp-not-inc	Exec-ma
2	Male	38	White	Divorced	HS-grad	United-States	Private	Handler
3	Male	53	Black	Married-civ-spouse	11th	United-States	Private	Handler
4	Female	28	Black	Married-civ-spouse	Bachelors	Cuba	Private	Prof-spe
5	Female	37	White	Married-civ-spouse	Masters	United-States	Private	Exec-ma
6	Female	49	Black	Married-spouse-absent	9th	Jamaica	Private	Other-se
7	Male	52	White	Married-civ-spouse	HS-grad	United-States	Self-emp-not-inc	Exec-ma
8	Female	31	White	Never-married	Masters	United-States	Private	Prof-spe
9	Male	42	White	Married-civ-spouse	Bachelors	United-States	Private	Exec-ma

	age	ldl
count	30,162	30,162
mean	38.4379	5.2414
std	13.1347	2.7415
min	17	0.5007
25%	28	2.857
50%	37	5.2452
75%	47	7.6119
max	90	9.9994

Abbildung 4.1: Grafische Oberfläche Daten-Upload (strukturiert)

Bei unstrukturierten Daten (TXT-Format) wird der gesamte Inhalt als Rohtext dargestellt. Eine strukturierte Analyse oder automatische Erkennung sensibler Informationen erfolgt erst in einem späteren Schritt. Eine tabellarische Vorschau oder statistische Auswertung wie bei strukturierten Daten findet hier nicht statt.

4.5 Erkennung sensibler Daten

Wie im Abschnitt 3.8 beschrieben, erfolgt die Erkennung sensibler Daten in strukturierter Form mittels einem dreistufigen Ablauf. Benutzerdefinierte Labels wurden definiert, um eine effiziente und domänenspezifische Analyse durchzuführen. Im nächsten Schritt erfolgt eine Überprüfung mithilfe vordefinierter regulärer Ausdrücke, die zuverlässig bekannte Muster erkennen. Im letzten Schritt wird mit spaCy natürliche Sprache verarbeitet. Für jeden erkannten Treffer im dreistufigen Ablauf wird der Sensitivitätsscore erhöht, wodurch sich eine Einschätzung zur Sensitivität der Daten ableiten lässt. Anhand dieser Metrik werden die Identifikatoren und Quasi-Identifikatoren automatisch identifiziert. Weiter können Nutzer:innen das Resultat der automatischen Erkennung über die grafische Oberfläche (vgl. Abbildung 4.2) manuell anpassen und eigene Identifikatoren und sensible Attribute definieren. Die erkannten Eigenschaften werden strukturiert aufbereitet, sodass sie im weiteren Verlauf gezielt für die Anonymisierung verwendet werden können.

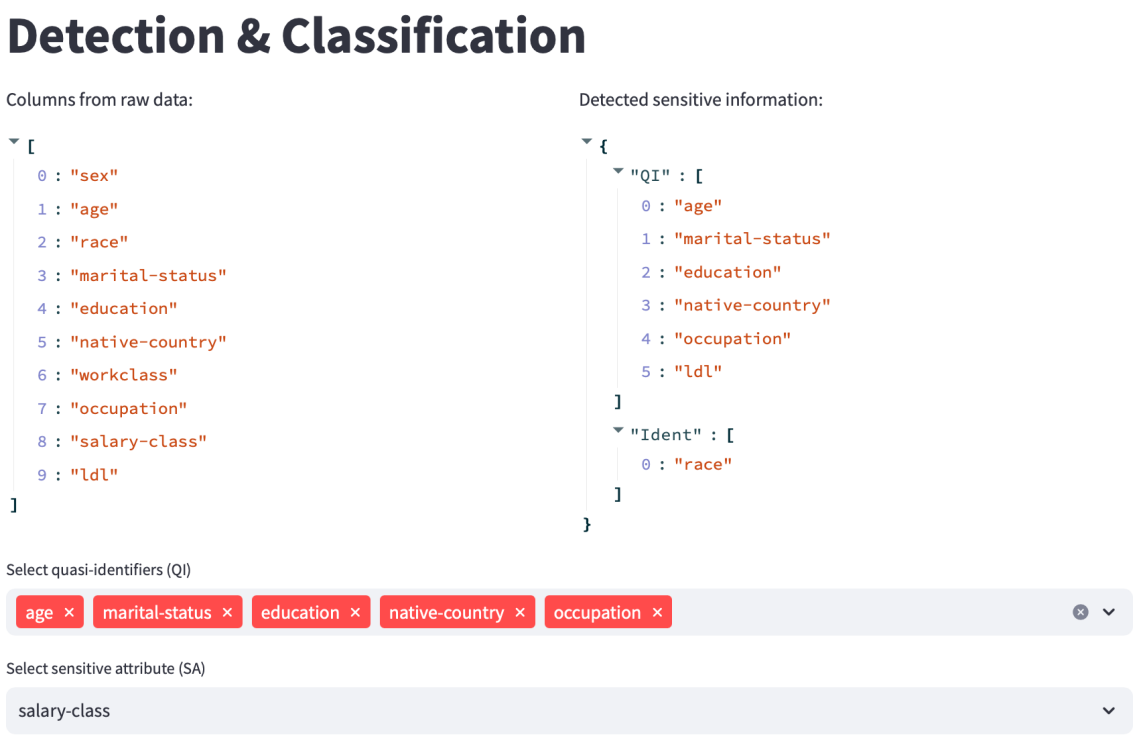


Abbildung 4.2: Grafische Oberfläche Erkennung sensibler Daten (strukturiert)

Bei der Erkennung sensibler Daten in unstrukturierter Form werden wiederum benutzerdefinierte Labels sowie reguläre Ausdrücke verwendet. Anstelle von spaCy wird hier Microsoft Presidio gebraucht, um mittels NER sensible Informationen kontextbasiert zu erkennen. Wie bei strukturierten Daten wurde ein Sensivitätsscore implementiert, um die Wörter zu klassifizieren. Die erkannten Begriffe werden visuell in der Oberfläche dargestellt (vgl. Abbildung 4.3). Die Treffer werden in einer Liste mit den Start- und End-Indizes sowie des Typs abgelegt, um sie für die Anonymisierung lokalisieren zu können.

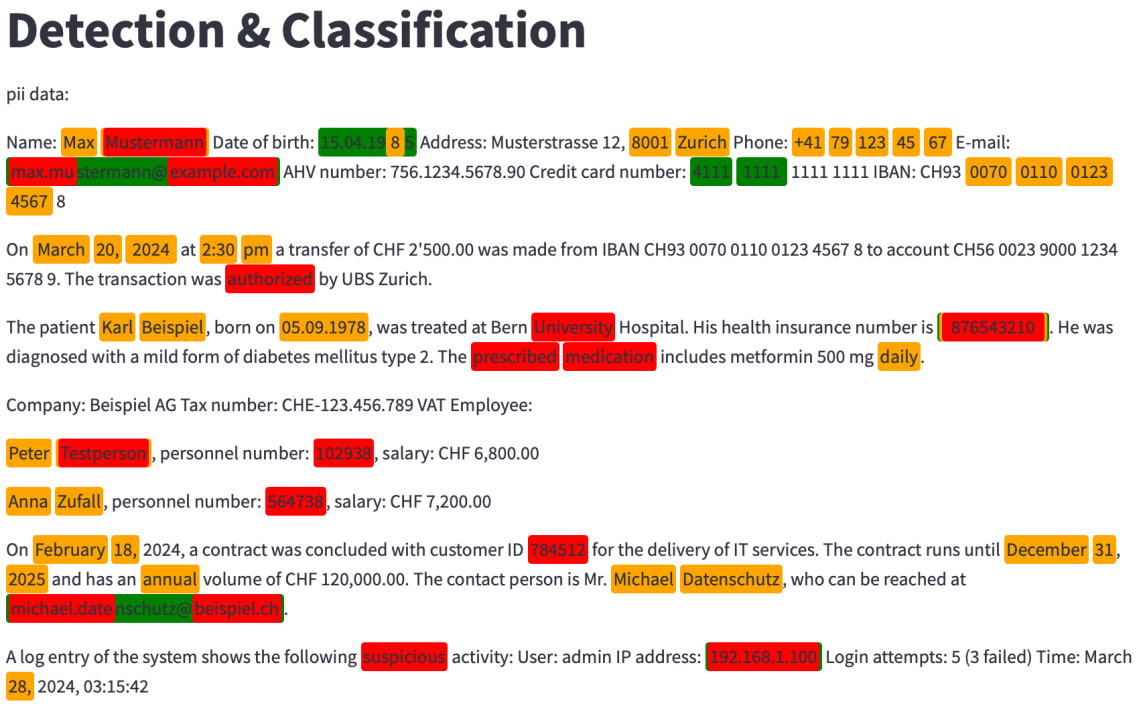


Abbildung 4.3: Grafische Oberfläche Erkennung sensibler Daten (unstrukturiert)

Insgesamt zeigte sich während der Entwicklung, dass die Erkennung bei strukturierten als auch bei unstrukturierten Daten zuverlässig funktionierte. In Einzelfällen konnten Fehlklassifikationen jeweils manuell korrigiert werden. Die Kombination aus regel- und NLP-basierten Verfahren erwies sich als praktikabel und gut anpassbar.

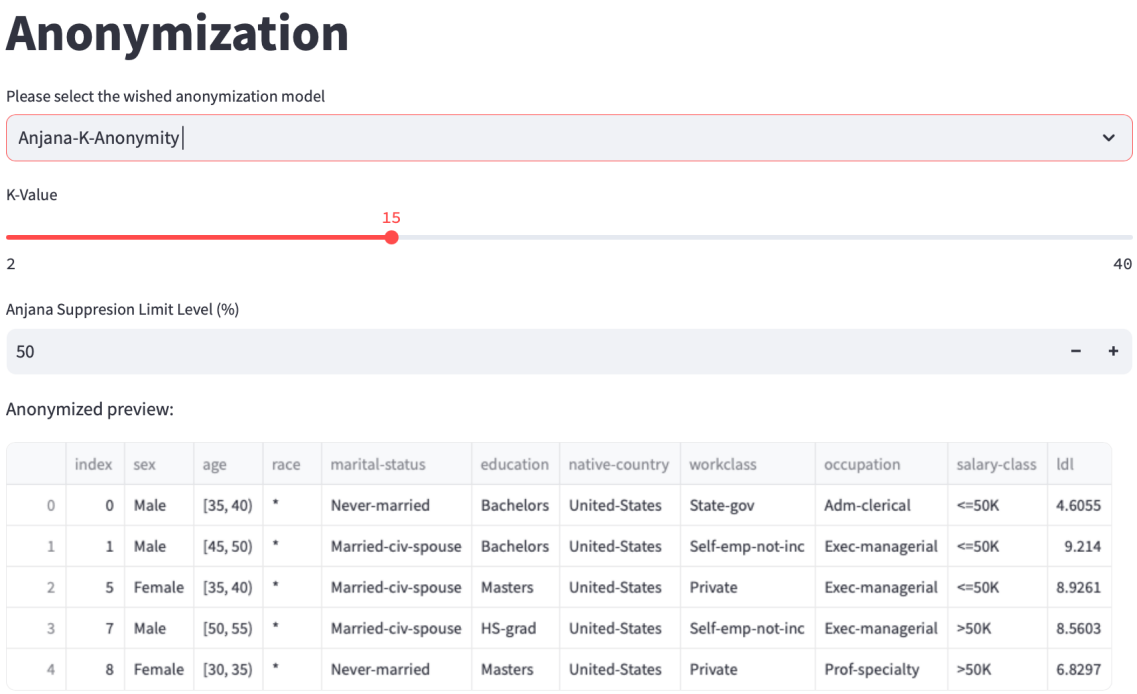
4.6 Anonymisierung

Zur Anonymisierung strukturierter Daten wurde die Bibliothek Anjana eingebunden (vgl. Abschnitt 3.3). Mit Anjana wurden die drei bewährten Anonymisierungsmodelle *k-Anonymität*, *ℓ-Diversität* und *t-Closeness* in das Webportal integriert. Nebst den Modellen von Anjana wurden anfangs eigene Implementationen von *k-Anonymität* und *ℓ-Diversität* zu Evaluationszwecken entwickelt. Die eigenen Implementationen dienten zur Überprüfung der grundlegenden Funktionsweise sowie als Referenz gegenüber den umfassenderen Modellen von Anjana. Obwohl im **??** bereits auf die implementierten (und mehr) Modelle eingegangen wird, werden sie hier kurz für die praktische Umsetzung zusammengefasst. Alle Modelle stehen in der grafischen Oberfläche den Nutzer:innen zur Auswahl:

k-Anonymität Die eigene k-Anonymität basiert auf einer simplen Generalisierung aller numerischen Werte. Dabei werden die Werte auf das nächstkleinere Vielfache eines frei wählbaren Faktors k gerundet. Dies führt zu einer Wertvergröberung, was das Wiedererkennen einzelner Werte erschwert. Die Funktion garantiert jedoch keine tatsächliche k-Anonymität im klassischen Sinne, da keine Äquivalenzgruppierung vorgenommen und keine Mindestgruppengrösse sichergestellt wird. Somit handelt es sich um ein illustratives Modell, nicht aber um eine vollständige k-Anonymität.

ℓ-Diversität Die eigene Implementation von ℓ-Diversität beschränkt sich auf eine einfache Gruppierung kategorialer Werte. Sobald mehr als fünf unterschiedliche Werte eines Attributs vorliegen, werden diese über eine Hash-Funktion zufällig auf fünf Gruppen verteilt. Diese Art der Verallgemeinerung reduziert den Informationsgehalt der Daten. Sie erfüllt jedoch nicht die Anforderungen der ℓ-Diversität. Die Gruppierung erfolgt nicht nach Quasi-Identifikatoren noch wird die Diversität innerhalb der Gruppen sichergestellt. Die Funktion dient lediglich als vereinfachtes Modell zur Illustration und nicht als vollständiges ℓ-diverses Modell.

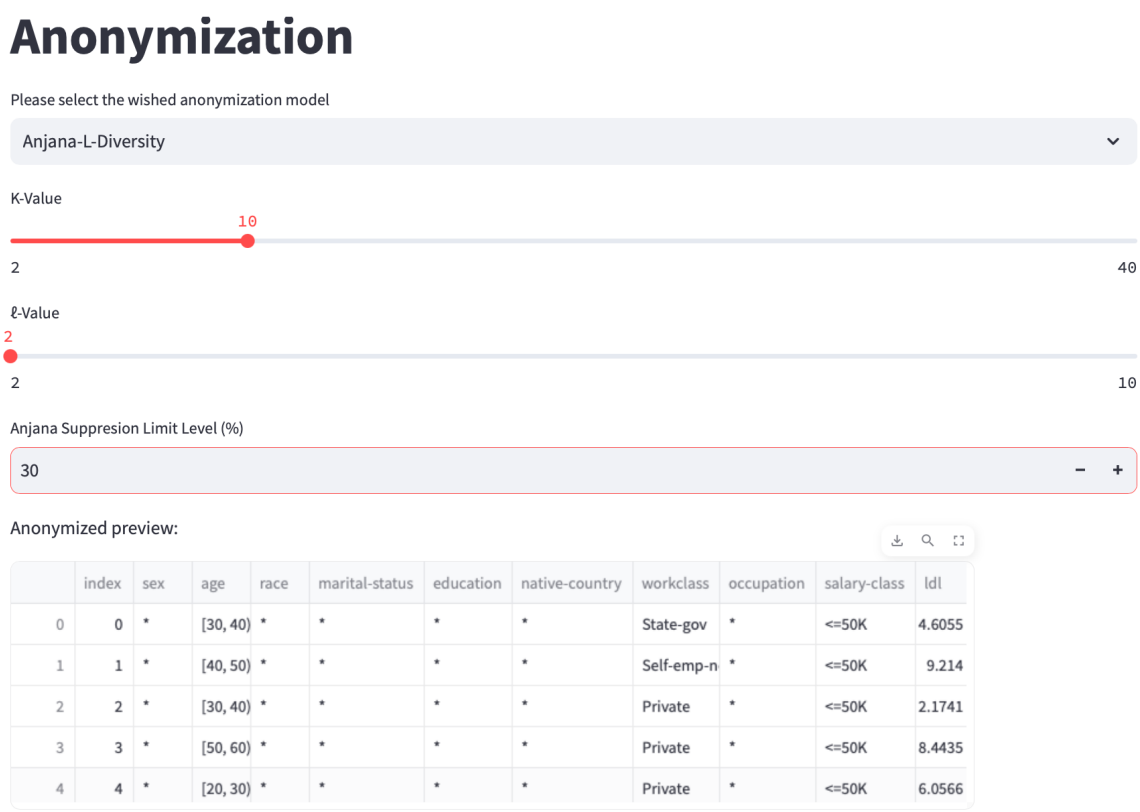
Anjana-k-Anonymität Die Implementierung der vollständigen k-Anonymität mit Anjana erfolgt unter Verwendung von fünf Parametern: `ident` (Identifikatoren), `qi` (Quasi-Identifikatoren), `k_value` (gewünschter k-Wert), `supp_level` (Unterdrückungslevel) und `hierarchies` (allfällige Generalisierungshierarchien). Diese k-Anonymität erzeugt Äquivalenzklassen mit mindestens k identischen Kombinationen der Quasi-Identifikatoren. Ist dies nicht realisierbar, werden einzelne Werte unter Berücksichtigung des Unterdrückungslevels entfernt. Die hierarchischen Strukturen dienen der gezielten Generalisierung (z.B. genaues Alter zu Wertebereich), um Daten besser gruppieren und den Informationsverlust steuern zu können. Auf dem Webportal werden die Identifikatoren, sowie Hierarchien bereits im Erkennungsschritt definiert. In der Abbildung 4.4 sieht man die Konfiguration des Anonymisierungsmodells sowie einer Vorschau der anonymisierten Daten.



	index	sex	age	race	marital-status	education	native-country	workclass	occupation	salary-class	ldl
0	0	Male	[35, 40)	*	Never-married	Bachelors	United-States	State-gov	Adm-clerical	<=50K	4.6055
1	1	Male	[45, 50)	*	Married-civ-spouse	Bachelors	United-States	Self-emp-not-inc	Exec-managerial	<=50K	9.214
2	5	Female	[35, 40)	*	Married-civ-spouse	Masters	United-States	Private	Exec-managerial	<=50K	8.9261
3	7	Male	[50, 55)	*	Married-civ-spouse	HS-grad	United-States	Self-emp-not-inc	Exec-managerial	>50K	8.5603
4	8	Female	[30, 35)	*	Never-married	Masters	United-States	Private	Prof-specialty	>50K	6.8297

Abbildung 4.4: Grafische Oberfläche Anonymisierung – Anjana-k-Anonymität

Anjana-ℓ-Diversität Die Anjana-Implementierung der vollwertigen ℓ-Diversität erweitert die k-Anonymität um ein zusätzliches Schutzkriterium. Zu den zuvor erwähnten Parametern kommt ein Diversitätsparameter `l_value` dazu. Die Anjana-ℓ-Diversität stellt sicher, dass in jeder Äquivalenzklasse mindestens ℓ unterschiedliche Ausprägungen des sensiblen Attributs enthalten sind. Durch die Angabe von Generalisierungshierarchien kann das Modell bessere Gruppierungen finden, bevor Werte unterdrückt werden müssen. In Abbildung 4.5 wird eine solche Anonymisierung auf dem Webportal gezeigt.



	index	sex	age	race	marital-status	education	native-country	workclass	occupation	salary-class	ldl
0	0	*	[30, 40)	*	*	*	*	State-gov	*	<=50K	4.6055
1	1	*	[40, 50)	*	*	*	*	Self-emp-n	*	<=50K	9.214
2	2	*	[30, 40)	*	*	*	*	Private	*	<=50K	2.1741
3	3	*	[50, 60)	*	*	*	*	Private	*	<=50K	8.4435
4	4	*	[20, 30)	*	*	*	*	Private	*	<=50K	6.0566

Abbildung 4.5: Grafische Oberfläche Anonymisierung – Anjana-ℓ-Diversität

Anjana-t-Closeness Das Anonymisierungsmodell t-Closeness von Anjana geht über die ℓ-Diversität hinaus, indem es die Verteilung der sensiblen Werte innerhalb einer Äquivalenzklasse mit der Gesamtverteilung im Datensatz vergleicht. Zusätzlich zu den bereits bekannten Parametern kommt hier der Schwellenwert `t_value` zum Einsatz. Dieser beschreibt den maximal erlaubten Unterschied zwischen den Verteilungen. Der Diversitätsparameter wird hier nicht mehr benötigt. Das Modell t-Closeness verwendet für die praktische Umsetzung geeignete Distanzmetriken wie beispielsweise die Earth Mover's Distance, um die Vertraulichkeit auch bei sensiblen Attributen mit dominanter Ausprägung zu schützen. Die Earth Mover's Distance misst den minimalen Aufwand, um eine Wahrscheinlichkeitsverteilung in eine andere umzuwandeln. Ein zu tiefer t-Wert kann das Gruppieren einschränken, während ein höherer Wert mehr Flexibilität bei geringerer Schutzwirkung erlaubt.

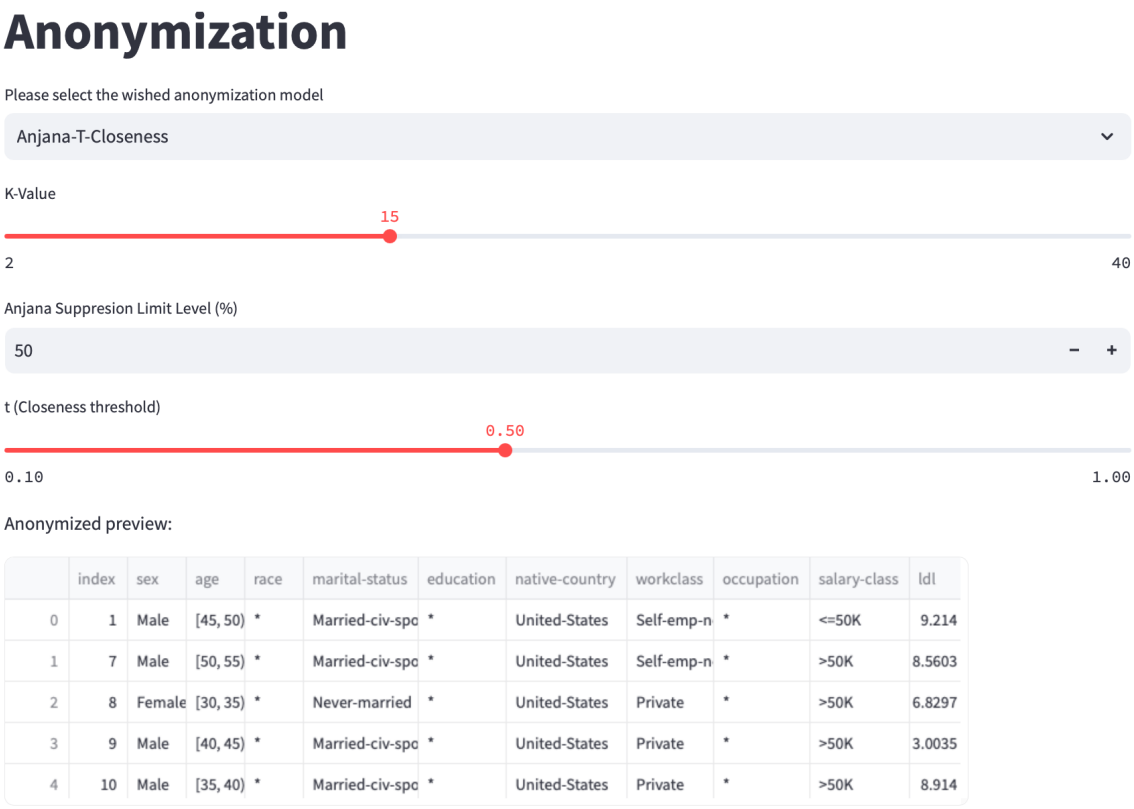


Abbildung 4.6: Grafische Oberfläche Anonymisierung – Anjana-t-Closeness

Bei Daten in unstrukturierter Form erfolgt die Anonymisierung über die Open-Source-Bibliothek Microsoft Presidio. Darin werden die erkannten Entitäten auf Wunsch pseudonymisiert oder maskiert (vgl. Abbildung 4.7). Da der Fokus auf der Anonymisierung strukturierter Daten lag, konnte diese Anonymisierung nicht weiter entwickelt werden.

Abbildung 4.7: Grafische Oberfläche Anonymisierung

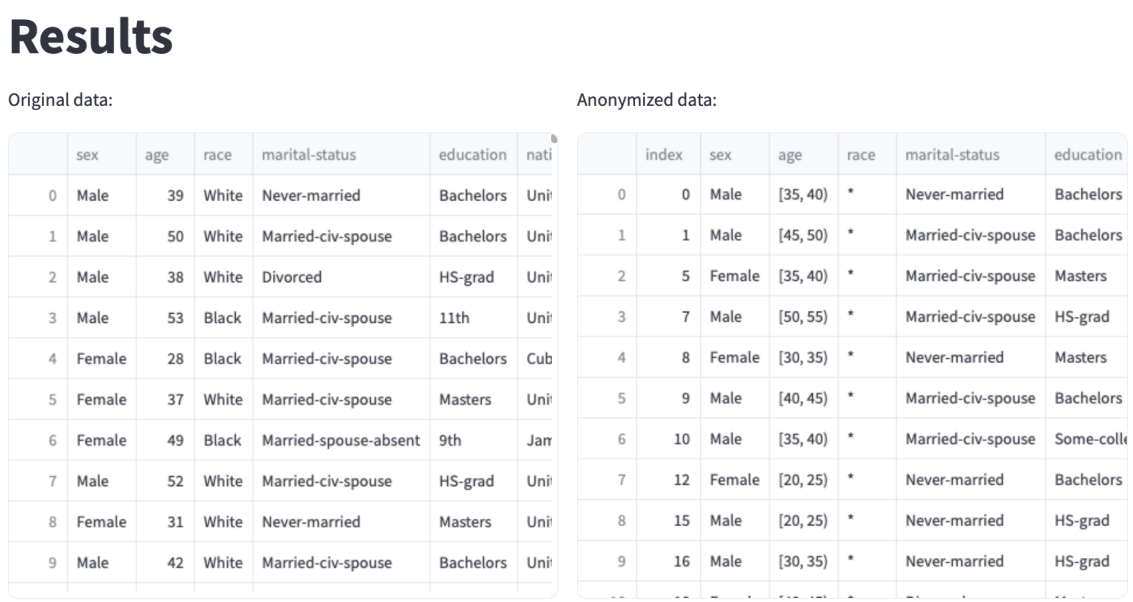
Die Auswirkungen der Anonymisierung lassen sich über unterschiedliche Metriken bewerten, welche im nächsten Abschnitt näher erläutert werden.

4.7 Visualisierung der Anonymisierungsergebnisse

Die erste Abbildung 4.8 zeigt einen direkten Vergleich der originalen und anonymisierten Daten. Die beiden Tabellen werden dabei nebeneinander dargestellt, damit Veränderungen auf den ersten Blick ersichtlich sind. Im Beispiel dieser k-Anonymität mit $k = 15$ wird ersichtlich, wie ein Index eingeführt und die Reihenfolge des Datensatzes durchmischt wird. Zudem wurde das Alter durch einen passenden Wertebereich ersetzt sowie die Rasse komplett unterdrückt.

Diese Visualisierung gibt einen intuitiven und transparenten Überblick der Anonymisierungsauswirkungen und dient als Grundlage für die anschliessende Bewertung durch spezifische Metriken.

Hinweis: Für die weiteren Visualisierungen wird das Beispiel mit der gleichen Anonymisierung verwendet.



	sex	age	race	marital-status	education	nati
0	Male	39	White	Never-married	Bachelors	Unit
1	Male	50	White	Married-civ-spouse	Bachelors	Unit
2	Male	38	White	Divorced	HS-grad	Unit
3	Male	53	Black	Married-civ-spouse	11th	Unit
4	Female	28	Black	Married-civ-spouse	Bachelors	Cub
5	Female	37	White	Married-civ-spouse	Masters	Unit
6	Female	49	Black	Married-spouse-absent	9th	Jam
7	Male	52	White	Married-civ-spouse	HS-grad	Unit
8	Female	31	White	Never-married	Masters	Unit
9	Male	42	White	Married-civ-spouse	Bachelors	Unit

	index	sex	age	race	marital-status	education
0	0	Male	[35, 40)	*	Never-married	Bachelors
1	1	Male	[45, 50)	*	Married-civ-spouse	Bachelors
2	5	Female	[35, 40)	*	Married-civ-spouse	Masters
3	7	Male	[50, 55)	*	Married-civ-spouse	HS-grad
4	8	Female	[30, 35)	*	Never-married	Masters
5	9	Male	[40, 45)	*	Married-civ-spouse	Bachelors
6	10	Male	[35, 40)	*	Married-civ-spouse	Some-colle
7	12	Female	[20, 25)	*	Never-married	Bachelors
8	15	Male	[20, 25)	*	Never-married	HS-grad
9	16	Male	[30, 35)	*	Never-married	HS-grad

Abbildung 4.8: Grafische Oberfläche Resultate Vorher/Nachher-Vergleich

Die Visualisierung in Abbildung 4.9 zeigt die Veränderung der Informationsentropie von Shannon [11] pro Attribut vor und nach der Anonymisierung. Die darunterliegende Berechnung ist in Gleichung 2.6 beschrieben.

Nach der Anonymisierung ist die Entropie in mehreren Attributen gesunken. Dies zeigt, dass gewisse Werte zusammengefasst oder generalisiert wurden, wodurch die Vorhersagbarkeit erhöht und somit das Risiko einer Re-Identifikation reduziert wurde.

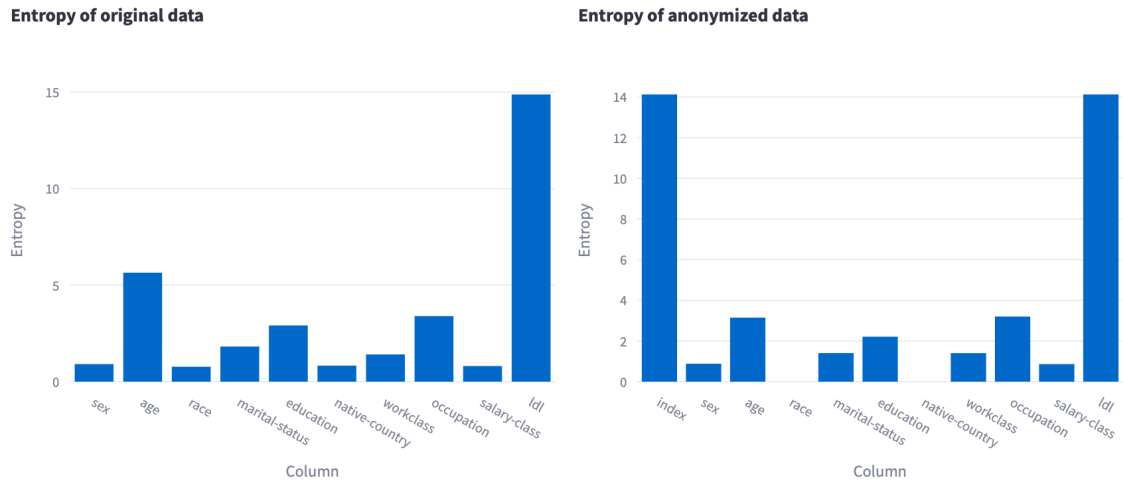


Abbildung 4.9: Grafische Oberfläche Resultate Entropie

Das linke Diagramm in Abbildung 4.10 zeigt den Informationsverlust pro Attribut, der durch die Differenz der Entropie vor und nach der Anonymisierung berechnet wird. Ein hoher Wert deutet auf einen stärkeren Informationsverlust hin. Durch Generalisierung oder Unterdrückung wurde ein Teil der ursprünglichen Variabilität der Daten entfernt.

Auf der rechten Seite wird übersichtlich dargestellt, welche Attribute unterdrückt worden sind. In diesem Beispiel wird lediglich das Attribut `race` vollständig unterdrückt. Diese Massnahme kann notwendig sein, wenn ein Attribut zu aussagekräftig ist oder nicht ausreichend generalisiert werden kann.

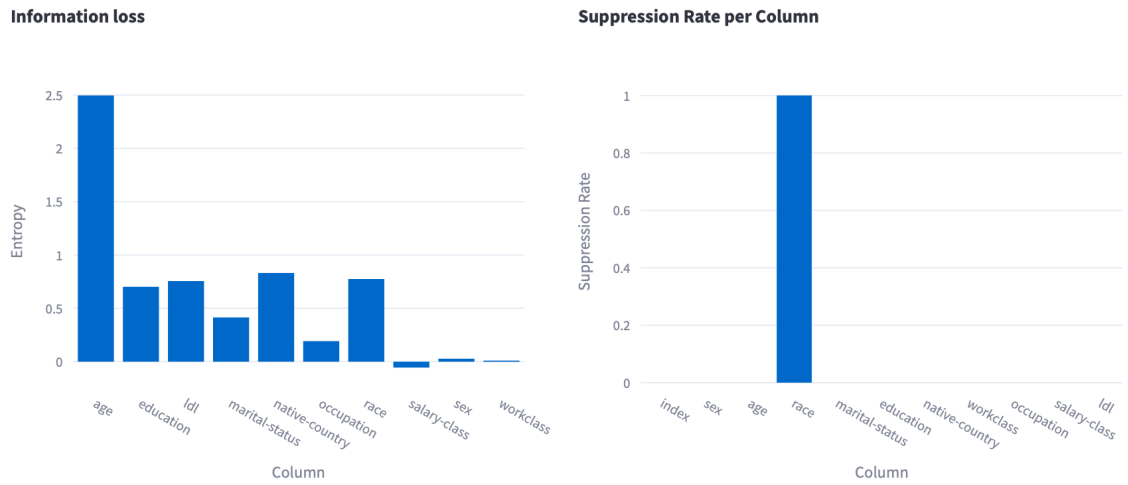


Abbildung 4.10: Grafische Oberfläche Resultate Informationsverlust und Unterdrückung

In den Ergebnissen des Webportals wird für jedes numerische Attribut automatisch ein Vorher-Nachher-Diagramm generiert, um die Veränderung der Werteverteilung durch die Anonymisierung zu visualisieren.

Die Abbildung 4.11 zeigt das Attribut `age` aus dem vorherigen Beispiel. Während in den Originaldaten konkrete Einzelwerte definiert sind, wurden diese im anonymisierten Datensatz zu Wertebereichen generalisiert. Es wird gut ersichtlich, dass die feinen Unterschiede verloren gehen und die Rückverfolgbarkeit reduziert wird. Jedoch ist hier auch ein gewisser Informationsverlust vorhanden.

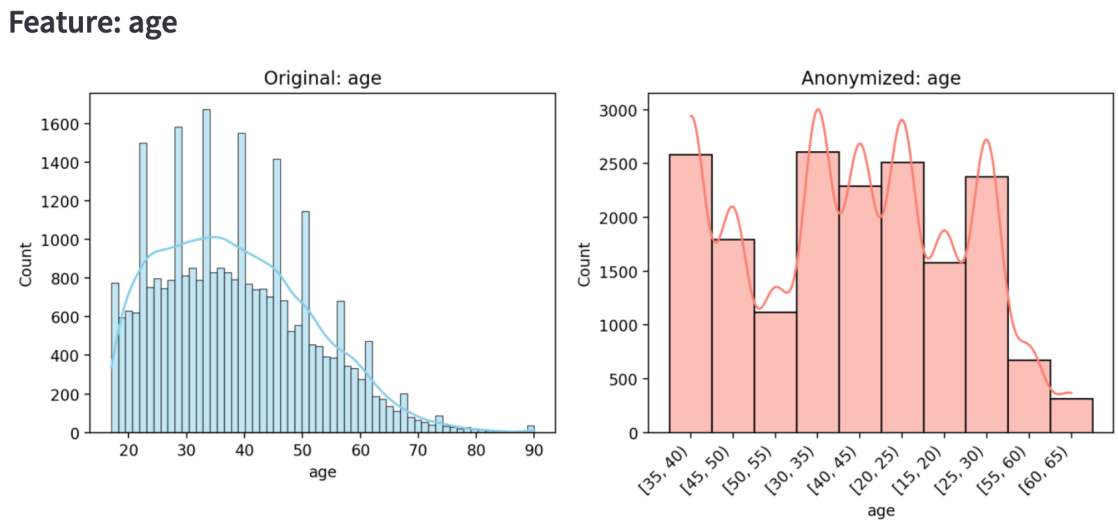


Abbildung 4.11: Grafische Oberfläche Resultate Alter

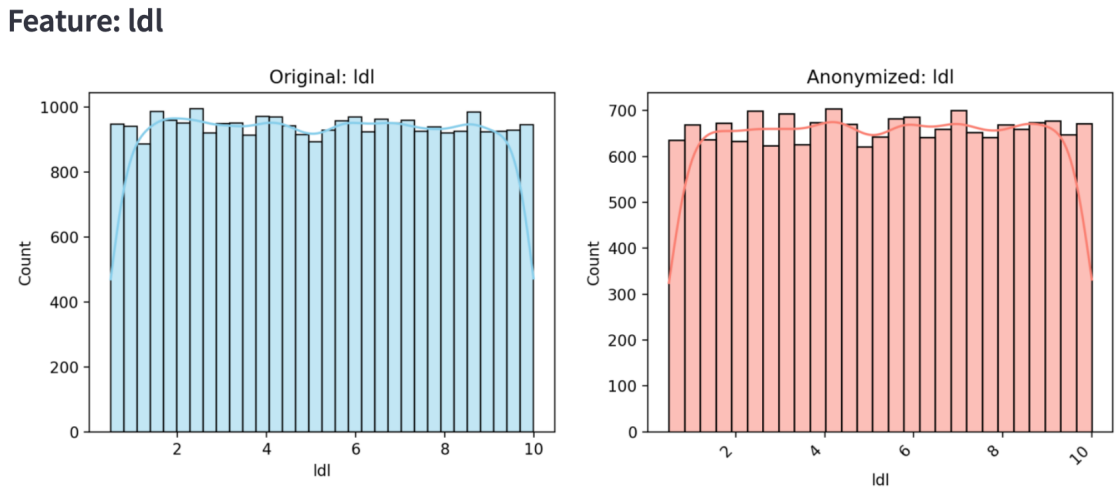


Abbildung 4.12: Grafische Oberfläche Resultate ldl

Eine weitere Abbildung 4.12 zeigt die Veränderung des numerischen Attributs `ldl` vor und nach der Anonymisierung. Durch die Balkenbreite ist ersichtlich, dass auch hier eine gewisse Generalisierung vorgenommen und die Aussagekraft der einzelnen Werte verringert wurde. Trotzdem bleibt die Form der Verteilung erkennbar, wodurch eine Analyse weiterhin möglich ist.

Die Abbildung 4.13 zeigt die Bewertung der Anonymität durch pyCANON vor und nach der Anonymisierung. Die Bewertung erfolgt anhand der in **??** erwähnten Metriken, die jeweils für die Rohdaten und die anonymisierten Daten berechnet werden.

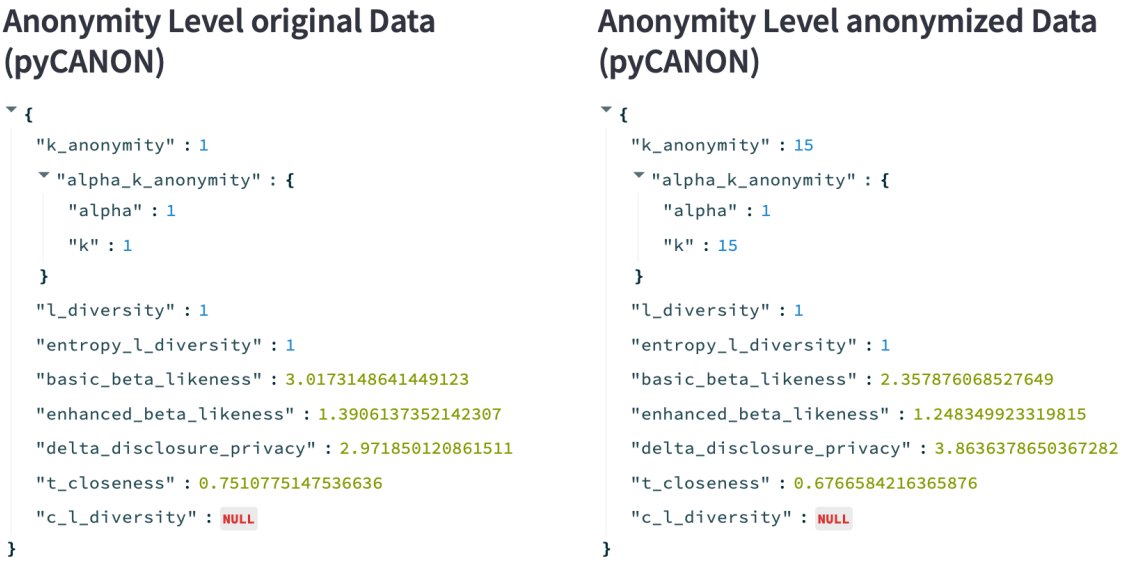


Abbildung 4.13: Grafische Oberfläche Resultate pyCANON Report

Bei diesem Beispiel wird die k-Anonymität von Anjana verwendet. Ein zentraler Unterschied zeigt sich beim k-Wert, der sich von $k = 1$ auf $k = 15$ erhöht hat. Dadurch wurde sichergestellt, dass jeder Datensatz mindestens mit 14 weiteren identisch ist. Dies ist eine signifikante Erhöhung des Datenschutzes.

Die ℓ-Diversität sowie die Entropie-ℓ-Diversität bleiben in beiden Versionen bei $\ell = 1$, was auf keine Verbesserung der Diversität innerhalb der Äquivalenzklassen hinweist. Dies ist verständlich, da nicht mit dem Modell ℓ-Diversität anonymisiert wurde.

Bei den β-Likeness-Metriken ist eine Veränderung sichtbar. Während die β-Likeness von etwa 3.02 auf 2.36 sinkt, zeigt sich bei der erweiterten Form ein leichter Rückgang von 1,39 auf etwa 1.25. Diese Rückgänge deuten auf eine Verringerung der maximalen Abweichung sensibler Verteilungen zwischen Gruppen hin. Dies ist ein positiver Effekt im Sinne des Datenschutzes.

Ein leichter Anstieg ist bei der δ-Disclosure Privacy zu verzeichnen. Diese Metrik quantifiziert, wie stark sich sensible Werte von der Gesamtverteilung abheben. Obwohl die k-Anonymität erhöht wurde, hat sich die Schutzwirkung gemessen an δ-Disclosure verschlechtert. Das heisst, dass die Verteilungen sensibler Werte innerhalb von Äquivalenzgruppen stärker von der Originalverteilung abweichen.

Die t-Closeness-Metrik verfolgt den Idealwert 0. In der Anonymisierung aus Abbildung 4.13 hat sich dieser Wert etwas verringert, also verbessert. Dies sagt aus, dass die Verteilung sensibler Attribute innerhalb von Äquivalenzgruppen der Gesamtverteilung stärker ähnelt.

4.8 Export

Der anonymisierte Datensatz kann auf der Resultat-Seite in unterschiedlichen Formaten exportiert werden. Unterstützt werden die Formate CSV, JSON und XML.

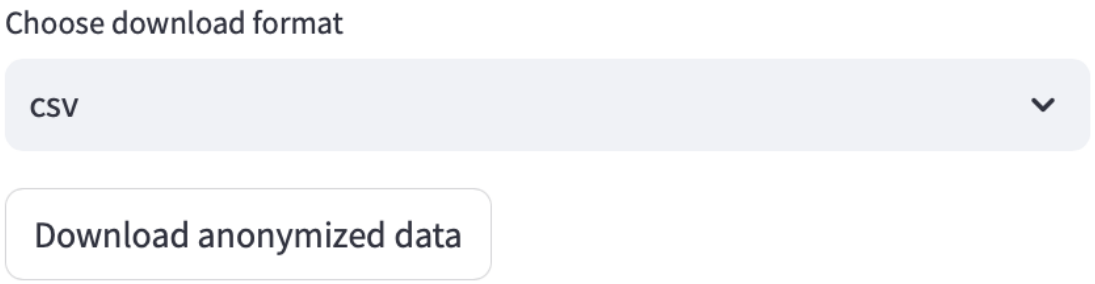


Abbildung 4.14: Grafische Oberfläche Resultate Export

4.9 Tests

Die Ergebnisse der in Abschnitt 3.4 definierten Tests werden in diesem Kapitel beschrieben.

4.9.1 Ergebnisse der Performanz-Tests

Zur Bewertung der Performanz der implementierten Anonymisierungsmodelle wurden Benchmark-Tests durchgeführt. Dabei wurde jeweils ein kleiner (3 Zeilen) sowie ein grosser (1 Million Zeilen) synthetischer Datensatz verwendet, um die Skalierbarkeit und Laufzeitunterschiede zu analysieren. Die Tests wurden automatisiert mit `pytest` durchgeführt. Alle Anonymisierungen wurden jeweils mit einem k-Wert von $k = 3$, einem ℓ-Wert von $\ell = 2$ und einem t-Wert von $t = 0.5$ durchgeführt.

Die Tabelle 4.5 zeigt einen Ausschnitt der relevantesten Ergebnisse der Tests in Mikrosekunden (μs). Für jedes getestete Modell sind die minimalen, maximalen, durchschnittlichen und die Median-Laufzeiten angegeben. Zusätzlich werden die errechneten Operationen pro Sekunde (OPS) sowie die Anzahl durchgeführter Runden dargestellt.

Test	Min	Max	Mean	Median	OPS	Rounds
`t_closeness_small`	38.75	91.13	41.97	41.67	23,826.13	7,486
`l_diversity_small`	90.92	179.58	99.90	100.33	10,010.50	4,733
`k_anonymity_small`	239.54	321.46	259.05	260.25	3,860.30	1,204
`t_closeness_large`	196,460.17	199,983.71	197,938.37	197,476.19	5.05	6
`k_anonymity_large`	207,117.21	208,296.96	207,730.40	207,947.29	4.81	5
`l_diversity_large`	462,156.63	512,144.79	484,557.14	473,489.63	2.06	5

Tabelle 4.5: Benchmark-Ergebnisse für verschiedene Anonymisierungsmodelle (Zeit in μs)

Aus diesen Ergebnissen lässt sich entnehmen, dass für kleine Datengrössen das Modell k-Anonymität am meisten Zeit benötigt. Beim grossen Datensatz hingegen braucht das Modell ℓ-Diversität mit Abstand am meisten Zeit.

4.9.2 Ergebnisse der Erkennung-Tests

Zur Überprüfung der Funktionalität Erkennungskomponente wurde ein gezielter Testfall durchgeführt. Dabei wurde ein synthetischer Datensatz erzeugt, der typische Attribute wie `Name`, `E-Mail-Adresse`, `Telefonnummer` und `Alter` beinhaltet.
Die Resultate zeigen, dass die Erkennung zuverlässig funktioniert:

- Alle identifizierenden Attribute wurden korrekt identifiziert und klassifiziert.
- Das Attribut `age` wurde als quasi-identifizierendes Attribut erkannt.

Es wurde zudem überprüft, ob die Rückgabestruktur der Funktion den Erwartungen entspricht. Die Tests verliefen erfolgreich und bestätigen die Korrektheit der Erkennung in der gegebenen Testumgebung. Der Vorteil dieses Unit-Tests liegt zusätzlich darin, dass bei zukünftigen Änderungen die Erkennungskomponente systematisch und automatisiert überprüft werden kann.

4.9.3 Ergebnisse der Anonymisierungs-Tests

Zur Validierung der Funktionalität der Anonymisierungskomponente wurden mehrere Unit-Tests auf einen synthetischen Datensatz durchgeführt. Dabei kamen alle in dieser Arbeit implementierten Anonymisierungsmodelle zum Einsatz.

Im Fokus standen folgende Aspekte:

- **Entropievergleich:** Bei allen Modellen wurde geprüft, ob die Entropie nach der Anonymisierung gegenüber dem Originaldatensatz reduziert oder gleich geblieben ist. Dieses Kriterium wurde in allen Fällen erfüllt, was auf einen erfolgreichen Informationsverlust im Sinne des Datenschutzes hinweist.
- **Strukturerhalt:** Die Struktur des ursprünglichen Datensatzes blieb nach der Anonymisierung erhalten. Dies ist für die Weiterverwendung der Daten essenziell.
- **Vollständigkeit der Daten:** Zur Sicherstellung der Datenqualität wurde überprüft, ob Werte fälschlicherweise `NULL` wurden.

4.10 Erfüllung der Anforderungen

Die in Tabelle 1.1 definierten Anforderungen wurden im Verlauf der Umsetzung regelmässig überprüft und bewertet. Die folgende Tabelle 4.6 dokumentiert den aktuellen Erfüllungsgrad aller Anforderungen. Ergänzend zur Bewertung wird ein kurzer Kommentar zur Erfüllung gegeben.

Id	Anforderung	Erfüllung	Kommentar
A1	Das Webportal soll Nutzer:innen die Möglichkeit bieten, strukturierte sowie unstrukturierte Daten hochladen zu können.	Erfüllt	Es können CSV-, JSON-, XML- und TXT-Dateien hochgeladen werden.
A2	Das Webportal soll sensible Daten automatisch erkennen und klassifizieren.	Erfüllt	Sensible Daten werden durch eine dreistufige Pipeline erkannt und klassifiziert.
A3	Das Webportal soll Nutzer:innen die Möglichkeit bieten, die automatisch klassifizierten Daten manuell anpassen zu können.	Erfüllt	Die vorgeschlagene Klassifizierung der Quasi-Identifikatoren und des sensiblen Attributs können manuell geändert werden.

Id	Anforderung	Erfüllung	Kommentar
A4	Das Webportal soll die k-Anonymität als Kernmethode zur Anonymisierung bereitstellen.	Erfüllt	Die k-Anonymität wurde als erste Methode in die modulare Architektur integriert und getestet.
A5	Die Nutzer:innen sollen die Anonymisierung durch Parameter wie den k-Wert konfigurieren können.	Erfüllt	Alle verfügbaren Parameter werden pro Anonymisierungsmodell dynamisch angezeigt können geändert werden.
A6	Das Webportal soll den Nutzer:innen Visualisierungen zu den Anonymisierungsauswirkungen bereitstellen.	Teilweise erfüllt	Visualisierungen werden den Nutzer:innen bereitgestellt, jedoch benötigt die Interpretation ein gewisses Vorwissen.
A7	Die anonymisierten Daten sollen in verschiedenen Formaten exportiert werden können.	Erfüllt	Die anonymisierten Daten können als XML, JSON oder CSV exportiert werden.
A8	Das Webportal soll eine intuitive und einfach bedienbare Benutzeroberfläche bieten.	Nicht erfüllt	Die Evaluation (z.B. Nutzer:innenstudie) konnte im Rahmen dieser Bachelorarbeit nicht durchgeführt werden.
A9	Das Webportal soll effizient mit grossen Datenmengen umgehen können.	Erfüllt	Benchmark-Tests haben gezeigt, dass auch mit grossen Daten, die Verarbeitungszeit sich im Rahmen befindet.
A10	Die Anonymisierung soll für eine gute Nutzer:innenerfahrung effizient durchgeführt werden.	Teilweise erfüllt	Die Anonymisierung wird durch effiziente Modelle von Anjana durchgeführt. Eine Nutzer:innenstudie für die Evaluierung ist hier von Vorteil.
A11	Das Webportal soll auf verschiedenen Endgeräten und Browsern funktionieren.	Erfüllt	Durch die Verwendung von Streamlit ist ist das Webportal auf allen gegenwärtigen Browsern verwendbar.
A12	Die Anonymisierungsmodelle sollen mit realitätsnahen Daten evaluiert werden.	Erfüllt	Für die Entwicklung des Webportals wurde stets ein realitätsnaher Datensatz verwendet.
A13	Das Webportal soll modular erweiterbar sein, um weitere Anonymisierungsmodelle einfach integrieren zu können.	Erfüllt	Weitere Anonymisierungsmodelle können mit einem geringen Aufwand in die modulare Architektur integriert werden.
A14	Die Erkennung sensibler Daten soll optional durch einen KI-gestützten Ansatz verfeinert werden.	Teilweise erfüllt	spaCy sowie Microsoft Presidio wurden integriert, jedoch müssen deren Genauigkeit genau evaluiert werden (vgl. Abschnitt 5.7).

Tabelle 4.6: Erfüllung der Anforderungen

5 Diskussion

In diesem Kapitel werden die Ergebnisse dieser Arbeit kritisch reflektiert, zentrale Erkenntnisse zusammengefasst sowie Limitationen und Verbesserungspotenziale identifiziert. Zudem wird ein Ausblick auf mögliche Weiterentwicklungen des Prototyps gegeben. Ziel ist es, sowohl den Nutzen als auch die Herausforderungen des entwickelten Webportals zur automatisierten Datenanonymisierung einzuordnen.

5.1 Bewertung der Zielerreichung

Die in Tabelle 1.1 definierten und in Tabelle 4.6 bewerteten Anforderungen wurden in weiten Teilen erfolgreich umgesetzt. Besonders hervorzuheben ist die effektive dreistufige Pipeline zur Erkennung sensibler Daten. Diese Pipeline kann verwendet werden, um entweder die Erkennung automatisch zu validieren, oder, um die Sensibilität der Erkennung zu steuern.

Ein zentrales Ziel war die Umsetzung eines dynamischen und transparenten Webportals zur Datenanonymisierung. Die Kernfunktionen wie das Hochladen von Daten, die Identifikation und Klassifikation sensibler Attribute sowie die Anwendung und Visualisierung unterschiedlicher Anonymisierungsmodelle konnten erfolgreich realisiert werden. Im Zuge dessen konnten alle Forschungsfragen aus **??** erfolgreich bearbeitet werden.

Funktionale Anforderungen wie die automatische Erkennung sensibler Daten, die dynamische Anonymisierung, der Export der anonymisierten Daten und weitere (A1-A5, A7, A9) wurden vollumfänglich erfüllt. Die Anforderung A6 (Visualisierung der Anonymisierungswirkungen) konnte teilweise erfüllt werden, da für die Interpretation der Visualisierungen ein spezifisches Vorwissen benötigt wird und in Richtung Verständlichkeit optimiert werden kann. Weiter wurde auch die funktionale Anforderungen A14 (Optionale Erkennung durch KI-gestützten Ansatz) nur teilweise erfüllt, da noch eine qualitative Evaluierung der verwendeten NLP-Modelle benötigt wird.

Die nicht-funktionalen Anforderungen A8 (Intuition der Benutzeroberfläche) und A10 (Effiziente Anonymisierung) wurden nicht erfüllt beziehungsweise teilweise erfüllt, da deren Erfüllung kaum objektiv bewertet werden kann. Eine Nutzer:innenstudie ist zukünftig erforderlich, um die Erfüllung dieser Anforderungen zu gewährleisten.

5.2 Balance zwischen Datenschutz und Datenqualität

Die Ergebnisse der Anonymisierung zeigen eine gelungene und transparente Balance zwischen Datenschutz und Nutzbarkeit der Daten. Zwar ist ein gewisser Informationsverlust unvermeidbar, dennoch bleibt ein hoher Anteil der Datenstruktur und Aussagekraft erhalten. Durch die verschiedenen Anonymisierungsmodelle und deren Parameter kann diese Balance individuell gesteuert werden. Die Visualisierungen aus Abschnitt 4.7 veranschaulichen die Auswirkungen der Anonymisierungen auf die Datenqualität und ermöglichen eine nachvollziehbare Abwägung zwischen Schutz und Nützlichkeit.

5.3 Stärken und Schwächen des Webportals

Stärken

- **Hybride Erkennung und Klassifikation:** Kombination von benutzerdefinierten Labels, RegEx und KI-gestützter Erkennung mit spaCy und Microsoft Presidio.
- **Modularer Aufbau:** Einfache Erweiterbarkeit für weitere Anonymisierungsmodelle oder Erkennungsmethoden.
- **Transparente Darstellung:** Nutzer:innen können die Prozesse mit Visualisierungen in jedem Schritt nachvollziehbar zur Laufzeit verfolgen.
- **Dynamik:** Das Portal passt sich flexibel an unterschiedliche Datenformate und Datensätze an.

Schwächen

- **Sprache der Modelle:** Das entwickelte Webportal wurde momentan nur mit englischsprachigen Daten und NER-Modellen evaluiert. Andere Sprachen könnten die Qualität und Funktionsweise einschränken.
- **Eingeschränkte Bewertung der Effizienz:** Ohne repräsentative Lasttests in einer produktiven Umgebung kann die Skalierbarkeit nur schwer eingeschätzt werden. Lokale Tests können je nach Auslastung und Gerät stark variieren.
- **Manuelle Nachkontrolle:** Die Qualität der Anonymisierung muss aktuell visuell durch Nutzer:innen geprüft werden. Eine systematische Bewertung wäre von Vorteil.

5.4 Limitationen der Arbeit

Diese Arbeit fokussierte sich auf die prototypische Umsetzung eines lokal ausführbaren Systems.

Folgende Limitationen sind zu beachten:

- Es wurde keine rechtskonforme Umsetzung nach DSGVO oder revDSG gemacht.
- Die Genauigkeit der verwendeten NLP-Modelle spaCy und Microsoft Presidio wurde nicht systematisch evaluiert.
- Es erfolgte keine Evaluation mit realen Endnutzer:innen oder Unternehmen.
- Die Anonymisierung unstrukturierter Daten konnte nur teilweise umgesetzt werden.

5.5 Risiken und Re-Identifizierbarkeit

Trotz der Anwendung bewährter Anonymisierungsmodelle wie k-Anonymität, ℓ-Diversität oder t-Closeness bleibt die Re-Identifizierbarkeit ein zentrales Risiko. Dieses Risiko wird durch den anhaltenden Trend zu Big Data zusätzlich verschärft. Denn durch die zunehmende Verfügbarkeit grosser, öffentlich oder kommerziell zugänglicher Datenquellen steigt die Gefahr, dass anonymisierte Datensätze mit externen Informationen verknüpft und so einzelne Personen doch wieder identifiziert werden können.

Ein konkretes Beispiel ist die Verknüpfung eines anonymisierten Gesundheitsdatensatzes mit öffentlich zugänglichen sozialen Netzwerken. Wenn beispielsweise Alter, Wohnregion und Krankheitsbild bekannt sind, könnten bereits wenige zusätzliche Merkmale ausreichen, um eine

Re-Identifizierung zu ermöglichen. Auch Daten aus Mobilfunkverbindungen, Kreditkartenkäufen oder Bewegungsmustern können verwendet werden, um scheinbar anonymisierte Daten zu deanonymisieren. Solche Angriffe sind bekannt unter der Bezeichnung Verknüpfungsangriffe.

Diese Bedrohung verdeutlicht die Notwendigkeit fortgeschrittener Anonymisierungsmodelle wie die differenzielle Privatsphäre. Im Gegensatz zu den implementierten Verfahren garantiert die differenzielle Privatsphäre mathematisch, dass das Hinzufügen von Rauschen die Gesamtausgabe der Analyse nur minimal beeinflusst. Dadurch wird eine Absicherung von Re-Identifikation erreicht, selbst bei grossem externen Wissen.

Ein weiteres Risiko ergibt sich aus der ausschliesslich englischsprachigen Implementierung und Evaluation. Werden anderssprachige Datensätze verarbeitet, kann die Qualität der Erkennung und Anonymisierung deutlich sinken. Besonders NER-Modelle sind sprachspezifisch, wodurch relevante Attribute in anderen Sprachen übersehen werden könnten. Eine mehrsprachige Erweiterung ist notwendig, um dieses Risiko zu minimieren.

5.6 Persönliche Reflexion

Die Entwicklung eines funktionierenden und vielseitigen Webportals zur automatisierten Datenanonymisierung stellte hohe Anforderungen an meine Selbstorganisation, Eigenverantwortung und Entscheidungsfindung. Die technische Umsetzung und Verschmelzung aller Komponenten erforderte ein tiefes Verständnis der zugrunde liegenden Konzepte. Aus diesem Grund musste ich viel Zeit in meine Literaturrecherche investieren.

Rückblickend war für mich vor allem die qualitative Erkennung sensibler Daten ein Schlüsselaspekt. Erst durch sie wird eine sinnvolle und wirksame Anonymisierung überhaupt möglich. Der tatsächliche Umfang war mir zu Beginn dieser Arbeit in dieser Tiefe nicht bewusst. Deshalb war es anfangs schwer einzuschätzen, welche Bereiche wie viel Zeit benötigen. Mit der Zeit und der Literaturrecherche zeigte sich, dass erwartete Arbeiten wie das Evaluieren oder Fine-Tuning von KI-Modellen nicht in den Rahmen dieser Arbeit passen.

Ein nicht zu unterschätzendes Hindernis waren die unterschiedlichen Datenformate. Unterschiedliche Strukturen erforderten flexible Lösungen und erschwerten die Erkennung und Anonymisierung. Mit diesem Hindernis konnte ich viel darüber lernen, wie dynamisch ein System aufgebaut sein muss, um realen Anforderungen gerecht zu werden.

Trotz begrenzter Ressourcen bin ich stolz darauf einen funktionsfähigen Prototypen mit einer Basis an Anonymisierungsmodellen realisiert zu haben. Das Projekt hat mir nicht nur viel technisches Wissen vermittelt, sondern auch meine Fähigkeit geschärft, komplexe Themen selbstständig und zielgerecht anzugehen.

5.7 Ausblick

Für zukünftige Arbeiten am Webportal lassen sich folgende Verbesserungspotenziale nennen:

- **Eigene Modelle trainieren:** Fine-Tuning von NLP-Modellen auf spezifische Datendomänen kann die Erkennungsleistung weiter verbessern.
- **Mehrsprachigkeit:** Erweiterung auf deutschsprachige (und weitere) NER-Modelle ist für reale Anwendungen unausweichlich.
- **Automatisierte Empfehlung:** Ein Algorithmus, der aus den erkannten sensiblen Daten passende Anonymisierungsmodelle und Parameter vorschlägt, wäre ein wertvoller Zusatz.
- **Qualitätskontrolle:** Eine automatisierte Prüfung der Qualität neben der visuellen Kontrolle würde die Sicherstellung der Anonymisierungsqualität verbessern.

- **Unstrukturierte Daten:** Die Verarbeitung unstrukturierter Daten sollte weiter ausgebaut werden. Die bestehende Erkennung und Anonymisierung ist zu optimieren. Für die Evaluation sollen Visualisierungen entwickelt werden, die die Qualität der Anonymisierung veranschaulichen.
- **Erweiterung von Anonymisierungsmodellen:** Neben klassischen Modellen könnte zukünftig auch die differenzielle Privatsphäre integriert werden, um mathematisch abgesicherte Datenschutzgarantien zu bieten.
- **Unterschiedliche Testdaten:** In der Entwicklung des Webportals wurde sich auf einen Datensatz konzentriert, um qualitative Aussagen zur Qualität der Erkennung und Anonymisierung machen zu können. Zukünftig muss die Anwendung mit weiteren Daten getestet werden.

Glossar

ℓ-Diversität ℓ-Diversität ist ein Datenschutzmodell, das die k-Anonymität erweitert, indem es verlangt, dass innerhalb jeder Äquivalenzklasse mindestens ℓ unterschiedliche Werte eines sensiblen Attributs vorkommen.

Big Data Big Data bezeichnet grosse, komplexe und schnell wachsende Datenmengen, die mit traditionellen Methoden nicht effizient verarbeitet werden können.

Differenzielle Privatsphäre Die differenzielle Privatsphäre ist ein mathematisches Datenschutzmodell, das sicherstellt, dass das Hinzufügen oder Entfernen eines einzelnen Datensatzes die Ausgabe einer Analyse nur minimal beeinflusst. Dies wird meist durch das gezielte Einfügen von Rauschen in die Daten oder Analyseergebnisse erreicht.

Earth Mover's Distance Die Earth Mover's Distance ist ein Distanzmass zur Messung der Unterschiedlichkeit zwischen zwei Wahrscheinlichkeitsverteilungen. Sie beschreibt den minimalen Aufwand, um eine Verteilung in eine andere umzuwandeln, wobei sowohl Menge als auch Distanz berücksichtigt werden.

Git Eine freie Software zur Versionskontrolle von Dateien.

GitHub Eine Plattform zur Versionskontrolle von Softwareprojekten im Team. Sie beruht auf Git.

Hash-Funktion Eine Hash-Funktion ist eine mathematische Funktion, die beliebig grosse Eingabedaten auf eine feste Ausgabelänge transformiert. Im Rahmen dieser Arbeit wurde sie verwendet, um Eingabedaten zu Ganzzahlen zu transformieren, um eine schnelle Idizierung zu ermöglichen.

k-Anonymität k-Anonymität ist ein Datenschutzmodell, das sicherstellt, dass jede Person in einem Datensatz mindestens mit k anderen identisch ist, um eine individuelle Identifikation zu verhindern. Dies wird durch Generalisierung oder Unterdrückung erreicht.

LaTeX LaTeX ist ein Textsatzsystem, das oftmals für wissenschaftliche Arbeiten verwendet wird. Es ermöglicht die einfache Handhabung von Referenzen, mathematischen Formeln und einer klaren Struktur.

Microsoft Presidio Presidio ist eine Open-Source-Bibliothek von Microsoft, die auf spaCy aufbaut. Presidio konzentriert sich speziell auf die Verarbeitung von personenbezogener Daten (PII). Im Kern benutzt diese Bibliothek regelbasierte und KI-geschützte Methoden, um sensible Daten automatisch zu identifizieren.

Quasi-Identifikatoren Quasi-Identifikatoren sind Attribute in einem Datensatz, die für sich genommen keine eindeutige Identifikation ermöglichen, jedoch in Kombination mit anderen Daten eine Identifikation ermöglichen können.

Repository Ein Speicherort für Code, der Versionskontrolle ermöglicht und oft für Softwareprojekte genutzt wird.

spaCy spaCy ist eine Open-Source-Bibliothek für NLP in Python. Unter anderen wird die in dieser Arbeit benötigte NER bereitgestellt.

t-Closeness t-Closeness ist ein Datenschutzmodell, das fordert, dass die Verteilung sensibler Attribute innerhalb jeder Äquivalenzklasse höchstens um einen Wert t von der Verteilung in der Gesamtpopulation abweicht. Dadurch wird verhindert, dass durch eine stark abweichende Verteilung Rückschlüsse auf sensible Werte gezogen werden können.

Verknüpfungsangriffe Verknüpfungsangriffe (engl. Linking Attacks) sind eine Technik, bei der anonymisierte Datensätze durch Kombination mit anderen Datenquellen re-identifiziert werden.

Akronyme

BERT Bidirectional Encoder Representations from Transformers.

BSc Bachelor of Science.

CRF Conditional Random Fields.

CSV Comma-Separated Values.

DSGVO Datenschutz-Grundverordnung der Europäischen Union.

ECTS European Credit Transfer and Accumulation System.

HSLU Hochschule Luzern.

IBAN International Bank Account Number.

JSON JavaScript Object Notation.

KI Künstliche Intelligenz.

KL-Divergenz Kullback-Leibler-Divergenz.

NER Named Entity Recognition.

NLP Natural Language Processing.

PII Personally Identifiable Information.

PoC Proof of Concept.

RegEx Regulärer Ausdruck (engl. Regular Expression).

revDSG revidiertes Datenschutzgesetz der Schweiz.

XML Extensible Markup Language.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] M. T. Islam und B. Khan, „Md. Toriqul Islam https://orcid.org/0000-0003-3004-770X Transparency International Bangladesh (TIB), Bangladesh Big Data and Analytics: Prospects, Challenges, and the Way Forward," in *Encyclopedia of Information Science and Technology*, 6. Aufl. IGI Global, Mai 2022, S. 1–30, ISBN: ISBN13: 9781668473665. DOI: 10.4018/978-1-6684-7366-5.ch048.

[2] V. Ciriani, S. De Capitani Di Vimercati, S. Foresti und P. Samarati, „Theory of privacy and anonymity," in *Algorithms and Theory of Computation Handbook: Special Topics and Techniques*, 2. Aufl. Chapman & Hall/CRC, 2010, S. 18-1–18-35, ISBN: 9781584888208.

[3] „Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates," Nr. 2016/679, 2016, Datenschutz-Grundverordnung, Art. 25, Amtsblatt der EU L 119, S. 1–88. Adresse: https://dsgvo-gesetz.de (besucht am 14. 02. 2025).

[4] „Bundesgesetz über den Datenschutz (Datenschutzgesetz, DSG)," 2023, SR 235.1, Art. 5 & 6, in Kraft seit 1. September 2023. Adresse: https://www.fedlex.admin.ch/eli/cc/2022/491/de (besucht am 14. 02. 2025).

[5] H. Balzert, M. Schröder und C. Schäfer, *Wissenschaftliches Arbeiten - Ethik, Inhalt & Form wiss. Arbeiten, Handwerkszeug, Quellen, Projektmanagement, Präsentation, 3. Auflage*. PROF. BALZERT STIFTUNG, 2022. DOI: 10.18420/LB-WissArbeiten.

[6] C. Winter, V. Battis und O. Halvani, „Herausforderungen für die Anonymisierung von Daten," 6. Nov. 2019. DOI: 10.18420/inf2019_53.

[7] „Understanding the Confusion Matrix in Machine Learning," GeeksforGeeks. (), Adresse: https://www.geeksforgeeks.org/confusion-matrix-machine-learning (besucht am 18. 03. 2025).

[8] L. Derczynski, „Complementarity, F-score, and NLP Evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, S. 261–266.

[9] C. J. Van Rijsbergen, „Further experiments with hierarchic clustering in document retrieval," *Information Storage and Retrieval*, Jg. 10, Nr. 1, S. 1–14, 1974.

[10] „Classification: Accuracy, recall, precision, and related metrics — machine learning," Google for Developers. (), Adresse: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall (besucht am 18. 03. 2025).

[11] C. E. Shannon, „A mathematical theory of communication," *The Bell System Technical Journal*, Jg. 27, Nr. 3, S. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[12] S.-H. Kim, C. Jung und Y.-J. Lee, „An entropy-based analytic model for the privacy-preserving in open data," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, S. 3676–3684. DOI: 10.1109/BigData.2016.7841035.

[13] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[14] S. Kullback und R. A. Leibler, „On information and sufficiency," *The annals of mathematical statistics*, Jg. 22, Nr. 1, S. 79–86, 1951.

[15] N. Li, T. Li und S. Venkatasubramanian, „t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*, IEEE, 2006, S. 106–115.

[16] G. Kaur, „Usage of regular expressions in NLP," *International Journal of Research in Engineering and Technology IJERT*, Jg. 3, Nr. 01, S. 7, 2014.

[17] J. Friedl, *Mastering regular expressions*. Ö'Reilly Media, Inc.", 2006.

[18] C. Sutton, A. McCallum et al., „An introduction to conditional random fields," *Foundations and Trends® in Machine Learning*, Jg. 4, Nr. 4, S. 267–373, 2012.

[19] S. Naseer, M. M. Ghafoor, S. bin Khalid Alvi et al., „Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance.," *Pakistan Journal of Multidisciplinary Research*, Jg. 2, Nr. 2, S. 293–308, 2021.

[20] A. Friebely, *Analyzing the efficacy of microsoft presidio in identifying social security numbers in unstructured text*. Utica University, 2022.

[21] E. Shamsinejad, T. Banirostam, M. M. Pedram und A. M. Rahmani, „A Review of Anonymization Algorithms and Methods in Big Data," *Annals of Data Science*, Jg. 12, Nr. 1, S. 253–279, 1. Feb. 2025, ISSN: 2198-5812. DOI: 10.1007/s40745-024-00557-w. Adresse: https://doi.org/10.1007/s40745-024-00557-w.

[22] J. Sáinz-Pardo Díaz und Á. López García, „A Python library to check the level of anonymity of a dataset," *Scientific Data*, Jg. 9, Nr. 1, S. 785, 2022.

[23] P. Samarati und L. Sweeney, „Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[24] L. Sweeney, „k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, Jg. 10, Nr. 05, S. 557–570, 2002.

[25] R. C.-W. Wong, J. Li, A. W.-C. Fu und K. Wang, „(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, S. 754–759.

[26] A. Machanavajjhala, D. Kifer, J. Gehrke und M. Venkitasubramaniam, „l-diversity: Privacy beyond k-anonymity," *Acm transactions on knowledge discovery from data (tkdd)*, Jg. 1, Nr. 1, 3–es, 2007.

[27] J. Cao und P. Karras, „Publishing microdata with a robust privacy guarantee," *arXiv preprint arXiv:1208.0220*, 2012.

[28] A. Wood, M. Altman, A. Bembenek et al., „Differential privacy: A primer for a non-technical audience," *Vand. J. Ent. & Tech. L.*, Jg. 21, S. 209, 2018.

[29] C. Dwork, A. Roth et al., „The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, Jg. 9, Nr. 3–4, S. 211–407, 2014.

[30] Six Sigma Black Belt, *Paarweiser Vergleich Nutzwertanalyse incl. Excel Vorlage*, Zuletzt abgerufen am 26. Februar 2025. Adresse: https://www.sixsigmablackbelt.de/paarweiser-vergleich/.

[31] H. F. Binner, *Methoden-Baukasten für ganzheitliches Prozessmanagement, Systematische Problemlösungen zur Organisationsentwicklung und -gestaltung*, 1. Aufl. Springer Gabler Wiesbaden, 19. Aug. 2015, S. XIV, 246, ISBN: 978-3-658-08408-0. DOI: 10.1007/978-3-658-08409-7.

[32] J. Sáinz-Pardo Díaz und Á. López García, „An Open Source Python Library for Anonymizing Sensitive Data," *Scientific data*, Jg. 11, Nr. 1, S. 1289, 2024.

[33] „Home - Microsoft Presidio.“ (), Adresse: https://microsoft.github.io/presidio/analyzer/ (besucht am 30. 03. 2025).

[34] „Separation of concerns (SoC),“ GeeksforGeeks. Section: Software Development. (), Adresse: https://www.geeksforgeeks.org/separation-of-concerns-soc/ (besucht am 10. 04. 2025).

[35] „Top programming languages 2024 - IEEE spectrum.“ (), Adresse: https://spectrum.ieee.org/top-programming-languages-2024 (besucht am 16. 03. 2025).

[36] „Top 10 programming languages for 2025,“ GeeksforGeeks. Section: GBlog. (), Adresse: https://www.geeksforgeeks.org/top-10-programming-languages-for-2025/ (besucht am 30. 03. 2025).

[37] K. Kumar, *Programming Languages: A Survey*. 27. Nov. 2019.

[38] Y. Akkem, B. Kumar und A. Varanasi, „Streamlit application for advanced ensemble learning methods in crop recommendation systems–a review and implementation,“ *Indian J Sci Technol*, Jg. 16, Nr. 48, S. 4688–4702, 2023.

[39] M. Khorasani, M. Abdou und J. H. Fernández, *Web Application Development with Streamlit, Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework* (Professional and Applied Computing, Apress Access Books, Professional and Applied Computing (R0)), 1. Aufl. Apress Berkeley, CA, 27. Aug. 2022, S. XXVII, 480, ISBN: 978-1-4842-8110-9. DOI: 10.1007/978-1-4842-8111-6.

[40] M. Knoll, „IT-Risikomanagement im Zeitalter der Digitalisierung,“ *HMD Praxis der Wirtschaftsinformatik*, Jg. 54, Nr. 1, S. 4–20, 1. Feb. 2017, ISSN: 2198-2775. DOI: 10.1365/s40702-017-0287-4. Adresse: https://doi.org/10.1365/s40702-017-0287-4.

Anhang

A Installationsanleitung

Dieses Kapitel beschreibt die erforderlichen Schritte, um das Webportal lokal auf dem Gerät in Betrieb zu nehmen.

Installation

Voraussetzungen

Folgende Software wird vorausgesetzt:

- Python (Version 3.x)
- Pip
- Git (optional, falls das Repository geklont wird)

Schritte zur Installation

1. Repository klonen oder herunterladen:

```
git clone https://github.com/derungsp/data-anonymizer.git
cd data-anonymizer
```

2. Virtuelle Umgebung erstellen (optional, empfohlen):

```
python -m venv venv
source venv/bin/activate  # macOS/Linux
venv\Scripts\activate  # Windows
```

3. Abhängigkeiten installieren:

```
pip install -r requirements.txt
```

B Bedienungsanleitung

Dieses Kapitel beschreibt die grundlegende Bedienung des entwickelten Webportals zur automatisierten Datenanonymisierung.

1. Start der Anwendung

Nach der erfolgreichen Installation kann das Webportal über den folgenden Befehl gestartet werden:

```
streamlit run app/main.py
```

2. Daten hochladen

- Navigieren Sie zur Seite “Upload” oder klicken Sie auf “Start”.
- Wählen Sie eine Datei im CSV-, JSON-, XML- oder TXT-Format aus.
- Klicken Sie auf “Next”.

3. Sensibilität analysieren

- Nach dem Upload erfolgt die Analyse sensibler Daten automatisch.
- Optional: Bei Bedarf können die Quasi-Identifikatoren oder das sensible Attribut abgeändert werden.
- Sensible Inhalte werden farblich hervorgehoben (unstrukturierte Daten).
- Klicken Sie auf “Next”.

4. Anonymisierung durchführen

- Wählen Sie das gewünschte Anonymisierungsmodell aus der Liste aus.
- Optional: Konfigurieren Sie die Parameter des ausgewählten Modells.
- Klicken Sie auf “Next”.

5. Ergebnisse anzeigen und exportieren

- Nach Abschluss der Anonymisierung werden die Resultate grafisch aufbereitet:
  - Vorher-Nachher-Vergleiche
  - Entropie- und Verteilungsanalysen

  - Anonymitätsreport von pyCANON

- Die anonymisierten Daten können im gewünschten Format heruntergeladen werden.

6. Hinweise zur Nutzung

- Das System ist lokal ausführbar und verarbeitet keine Daten über externe Server.
- Die Erkennung basiert auf vordefinierten Regeln und Modellen. Eine Nachkontrolle durch Nutzer:innen bzw. Entwickler:innen ist empfohlen.
- Bei sensiblen Daten sollten Nutzer:innen zusätzlich rechtliche Rahmenbedingungen beachten (z.B. DSGVO oder revDSG). Das Webportal verspricht keine Datenschutz-Konformität!

C Kopie der Aufgabenstellung

Modul:	Dept I BAA FS25

Titel:	Entwicklung eines funktionsfähigen Webportals zur automatisierten und dynamischen Datenanonymisierung
Ausgangslage und Problemstellung:	Die Anforderungen an den Schutz personenbezogener Daten nehmen weltweit zu, getrieben durch Datenschutzgesetze und Notwendigkeit sicherer Datenverarbeitung. Datenanonymisierung ist ein entscheidender Ansatz, um Datenschutz zu gewährleisten, aber bestehende Lösungen sind oft schwer anpassbar und technisch anspruchsvoll. Ziel ist es, ein funktionsfähiges Webportal zu entwickeln, das eine automatisierte Anonymisierung durchführt. Das System soll sensible Daten erkennen, klassifizieren und diese anschliessend anonymisieren. Die Auswirkungen der Anonymisierung sollen durch verschiedene Visualisierungen (z.B. Vorher-Nachher-Vergleich, Heatmaps, Entropieanalysen) dargestellt werden.
Ziel der Arbeit und erwartete Resultate:	Die Arbeit soll ein funktionsfähiges Webportal hervorbringen, das: • eine automatisierte Erkennung und Klassifizierung sensibler Daten ermöglicht. • geeignete Anonymisierungsmodelle basierend auf den Datencharakteristiken vorschlägt. • Anonymisierung auf Basis von K-Anonymität als Kernmethode implementiert. • die Datenschutzwirkung durch Visualisierungstechniken wie Risiko-Heatmaps, Histogramme und Entropieanalysen darstellt. • eine erweiterbare Architektur bietet, die zukünftig für andere Anonymisierungsmodelle erweitert werden kann. Das Portal soll mit realitätsnahen Kundendaten (z.B. aus dem E-Commerce- oder Gesundheits-Sektor) evaluiert werden.

Gewünschte Methoden, Vorgehen:	• Anforderungsanalyse und Konzeptentwicklung:
	  ○ Literaturrecherche zu bestehenden Anonymisierungsmodellen sowie zu Methoden der automatisierten Identifikation sensibler Daten
	  ○ Use-Case-Definition zur Identifikation der Kernanforderungen
	• Entwicklung des Webportals
	  ○ Kanban als Vorgehensmodell zur effizienten Steuerung der Aufgaben im kleinen Team
	  ○ Software-Engineering-Prinzipien für eine erweiterbare Architektur
	  ○ Backend: Implementierung der Anonymisierungsalgorithmen
	  ○ Frontend: Darstellung/Steuerung der Anonymisierungsmodelle
	• Implementierung der Anonymisierungsalgorithmen:
	  ○ Kernmethode: K-Anonymität
	  ○ Erweiterungsmöglichkeiten: Andere Anonymisierungsmodelle als optionale Verbesserungen

Abbildung C.1: Kopie der Aufgabenstellung Teil 1

	○ Dynamische Parameter (z.B. Schwellenwerte für K) • Evaluation und Visualisierung ○ Simulation realistischer Szenarien (z.B. mit Kundenstammdaten eines Unternehmens) ○ Qualitätsbewertung der anonymisierten Daten mittels Entropieanalyse und Informationsverlust-Messungen und Vergleich mit anderen Tools ○ Visualisierung der Resultate (Vorher-Nachher, Heatmaps, Histogramme…)
Kreativität, Methoden, Innovation:	• Interaktive Visualisierung: Nutzende sehen, wie die Anonymisierung die Daten verändert, z. B. durch Heatmaps oder Filteroptionen. • Modular erweiterbare Architektur: Integration anderer Anonymisierungsmodellen mit minimalem Entwicklungsaufwand. • Optionale KI-Integration: Automatische Identifikation sensibler Daten mittels Machine Learning-Techniken wie Named Entity Recognition (NER)
Sonstige Bemerkungen:	Die optionale KI-Integration wird als innovativer Bestandteil begrüsst, ist jedoch keine zwingende Voraussetzung.

Projektteam

Student 1:	Pascal Derungs
Betreuer:in:	Radwan Eskhita

Auftraggeber

Firma:	HSLU
Ansprechperson:	Radwan Eskhita
Funktion:	Dozent
Strasse:	
PLZ/Ort:	
Telefon:	079 6657472
E-Mail:	radwan.eskhita@hslu.ch
Website:	

Version 13.06.2023 / bcl

Abbildung C.2: Kopie der Aufgabenstellung Teil 2

D Risikoanalyse

Identifikation der Risiken

Die folgende Tabelle zeigt die identifizierten Risiken dieser Arbeit. Diese Risikoanalyse dient der internen Orientierung im Projekt.

Id	Name	Beschreibung
R1	Qualität der Anonymisierung	Bei einer zu starken Anonymisierung droht Informationsverlust. Bei einer zu schwachen Anonymisierung könnte die Zuordnung zu einer Person immer noch gewährleistet sein.
R2	Projektumfang	Durch neue Erkenntnisse oder Anforderungen könnte sich der Projektumfang als zu gross entpuppen.
R3	Wissen (Methoden zur Erkennung sensibler Daten)	Der Autor verfügt über begrenztes Wissen in Sachen Methoden zur Erkennung sensibler Daten.
R4	Wissen (KI-gestützte Erkennung)	Der Autor verfügt über begrenztes Wissen in der Benutzung von KI-gestützten Methoden.
R5	Wissen (Anonymisierungsmodelle)	Der Autor verfügt über begrenztes Wissen in den Anforderungen und Anwendungen von Anonymisierungsmodellen.
R6	Technische Komplexität	Der Gesamtkontext (Erkennung, Anonymisierung und Visualisierung) könnte technisch zu komplex sein und den Zeitplan überschreiten.
R7	Datenqualität	Durch unpassende Testdaten könnten die Ergebnisse verfälscht sein.
R8	Leistungsprobleme	Durch Auswahl von ineffizienten Algorithmen könnten sich Leistungsprobleme ergeben.
R9	Krankheit	Krankheitsbedingte Ausfälle könnten Verzögerungen im Zeitplan verursachen.
R10	Benutzerfreundlichkeit	Der Autor verfügt keine tiefe UX/UI-Erfahrung und so besteht das Risiko, dass die Benutzeroberfläche des Webportals schwer verständlich wird.

Tabelle D.1: Identifikation der Risiken

Bewertung der Risiken

Wie bereits in Abschnitt 1.8 erwähnt, ist die Bewertung der Risiken subjektiv. Die Bewertung wird mit einer 5x5 Risikomatrix und numerischen Werten vorgenommen [40, S. 15]. Pro Risiko wird für das Schadensausmass sowie für die Eintrittswahrscheinlichkeit jeweils ein Wert zwischen 1 und 5 gewählt.

Die Bewertung folgt aus dem Produkt der beiden Zahlen:

- 0-5: Geringes Risiko
- 6-9: Mittleres Risiko
- $\geq$ 10: Hohes Risiko (Massnahmen erforderlich)

Id	**Schadensausmass (S)** (1-5)	**Eintrittswahrscheinlichkeit (E)** (1-5)	**Bewertung** (S $\times$ E)
R1	5	3	15
R2	4	3	12
R3	3	3	9
R4	3	2	6
R5	3	3	9
R6	4	4	16
R7	3	2	6
R8	4	3	12
R9	3	3	9
R10	3	3	9

Tabelle D.2: Risikomatrix vor den Minderungsmassnahmen

Massnahmen zur Risikominderung

Aus Risiken mit einer Bewertung ≥ 10 werden Massnahmen definiert, um das Schadensausmass oder die Eintrittswahrscheinlichkeit zu minimieren.

R1 - Qualität der Anonymisierung Eine vertiefte Literaturrecherche wird im Bereich der Datenanonymisierung durchgeführt. Weiter wird die Qualität der Anonymisierung an realitätsnahen Datensätzen evaluiert. Eintrittswahrscheinlichkeit $3 \rightarrow 2$

R2 - Projektumfang Eine klare Anforderungs- und Risikoanalyse wird zu Beginn des Projekts durchgeführt. Daraus wird ein Zeitplan mit Meilensteinen definiert. Eintrittswahrscheinlichkeit $3 \rightarrow 2$

R6 - Technische Komplexität Mit der ausgewählten Programmiersprache wird ein schneller Proof of Concept (PoC) erstellt, um alle grundlegenden Funktionen zu testen. Eintrittswahrscheinlichkeit $4 \rightarrow 2$

R8 - Leistungsprobleme Eine vertiefte Literaturrecherche wird im Bereich der Effizienz von Programmiersprachen, Technologien und Verfahren durchgeführt. Eintrittswahrscheinlichkeit $3 \rightarrow 2$

Aus den Minderungsmassnahmen von Tabelle D ergibt sich neu die folgende Matrix:

Id	**Schadensausmass (S)** (1-5)	**Eintrittswahrscheinlichkeit (E)** (1-5)	**Bewertung** (S × E)
R1	5	2	10
R2	4	2	8
R6	4	2	8
R8	4	2	8

Tabelle D.3: Risikomatrix nach den Minderungsmassnahmen

E Vorgehensmodell

In Abbildung E.1 wird das Kanban-Board für die Abwicklung dieser Arbeit gezeigt.

Kanban-Board

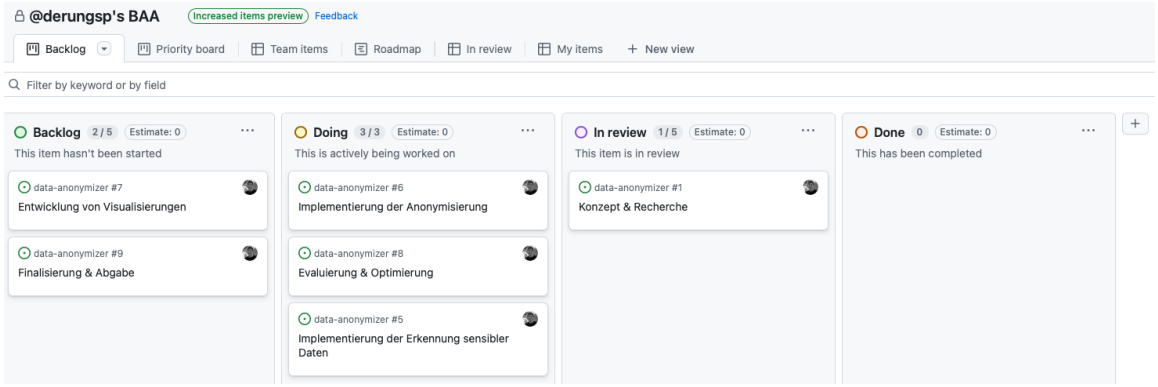


Abbildung E.1: Kanban-Board (Stand Februar)

Meilensteine

In Abbildung E.2 werden die definierten Meilensteine gezeigt, die für die grobe Zeitplanung dieser Arbeit erstellt wurden.

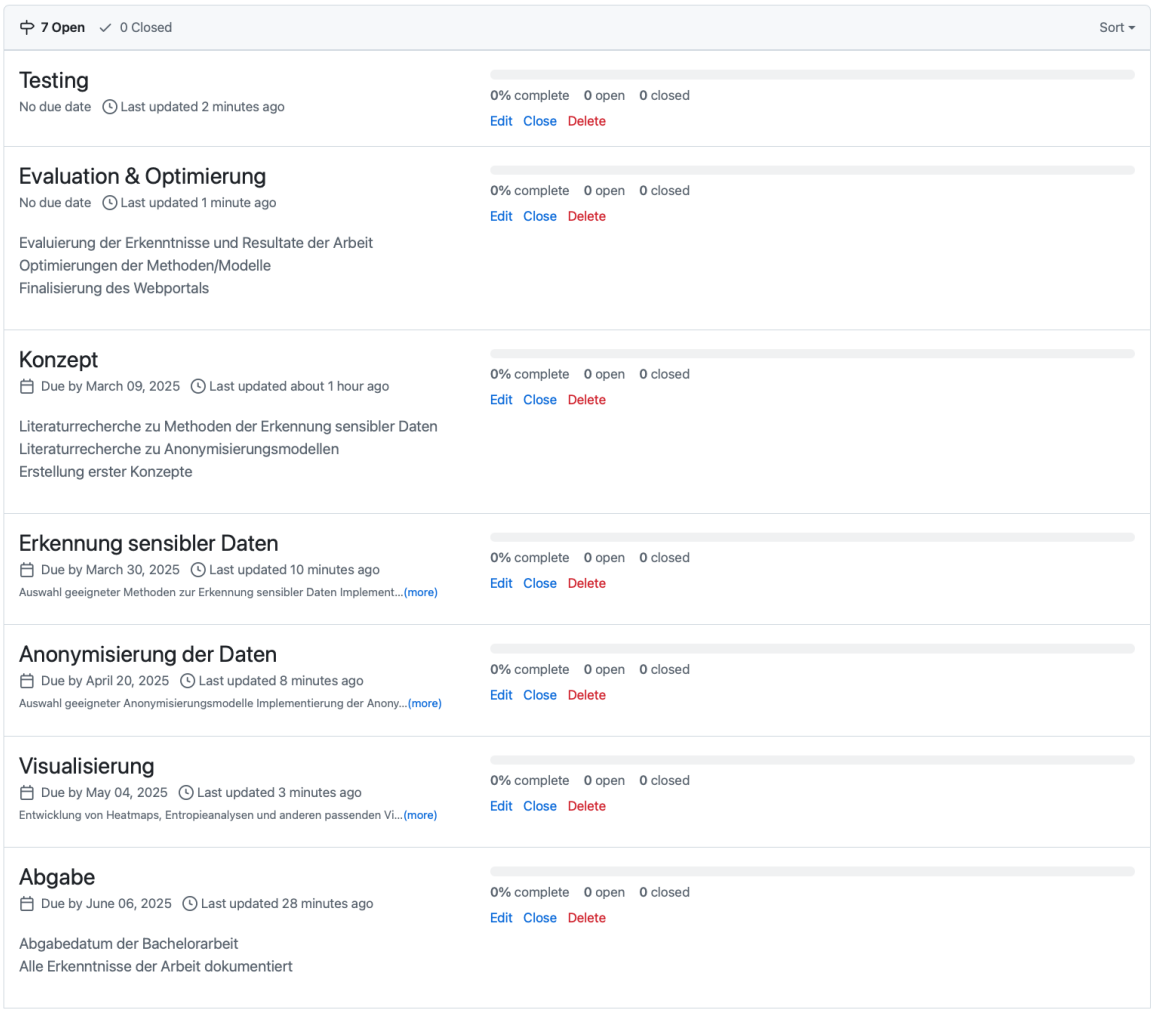


Abbildung E.2: Übersicht der Meilensteine (Stand Februar)

F Arbeitsjournal

Einleitung

Das Arbeitsjournal in Tabelle F dokumentiert die Fortschritte der Bachelorarbeit, strukturiert nach den einzelnen Arbeitstagen. Es enthält eine kurze Beschreibung der Tätigkeiten sowie eine Protokollierung des Aufwands in Stunden. Dies ermöglicht einen groben Soll-Ist-Vergleich über den Fortschritt der Arbeit im Vergleich zur insgesamt verfügbaren Bearbeitungszeit von 360 Stunden.

Übersicht der Arbeitspakete

Datum	**Aufgabe**	**Aufwand**
14.02.2025	Literaturrecherche und Vorbereitung für Kick-off-Pitch	6h
17.02.2025	Erstellung / Konfiguration LaTeX-Dokumentation	4h
18.02.2025	Einrichtung Zotero, Erstellung Bibliographie, Literaturrecherche im Bereich Datenanonymisierung (+ revDSG, DSGVO)	6h
19.02.2025	Kick-off Termin, Spezifizierung der Aufgabenstellung	3h
20.02.2025	Literaturrecherche zur Erkennung sensibler Daten, Überarbeitung der Aufgabenstellung	3h
23.02.2025	Literaturrecherche zur Erkennung sensibler Daten	2h
24.02.2025	Finalisierung Aufgabenstellung	2h
25.02.2025	Anforderungsanalyse / Use-Case-Definition	3h
26.02.2025	Anforderungsanalyse / Use-Case-Definition und Risikoanalyse	6h
27.02.2025	Präferenzanalyse / Nutzwertanalyse Tech-Stack Backend und Frontend	6h
01.03.2025	Präferenzanalyse / Nutzwertanalyse Tech-Stack Backend und Frontend, Use-Case-Diagramm erstellt und Anforderungen überarbeitet	5h
02.03.2025	Use-Case-Diagramm erstellt und Anforderungen überarbeitet	5h
03.03.2025	Risikoanalyse abschliessen und Matrizen erstellen	6h
04.03.2025	Erstellung der Risikomatrizen abschliessen und Minderungsmassnahmen definieren, Meeting geplant	5h

Datum	Aufgabe	Aufwand
06.03.2025	Vertiefung in Erkennungsmodelle und diverse Metriken (F1-Score, Entropie), Suche nach passenden, realitätsnahen Testdaten	3h
08.03.2025	Suche nach passenden, realitätsnahen Testdaten, Recherche nach Arbeiten und Literatur mit deutschsprachigen Modellen	4h
09.03.2025	Erstellung Proof of Concept der Anonymisierungspipeline	8h
10.03.2025	Vertiefung in Theorie von Anonymisierungsmodellen (k-Anonymität, ℓ-Diversität, Differential Privacy)	6h
11.03.2025	Nachführen der Dokumentation	4h
12.03.2025	Nachführen der Dokumentation, Paarweiser-Vergleich / Nutzwertanalyse untermauern	4h
15.03.2025	Nachführen der Dokumentation, Recherche zu NER (spaCy und Microsoft Presidio)	3h
16.03.2025	Erstellung PoC mit spaCy	4h
17.03.2025	Recherche Metriken zur Evaluation von NLP-Modellen	3h
18.03.2025	Recherche Metriken zur Evaluation von NLP-Modellen, Nachführen der Dokumentation, Umstellung nach der LaTeX-Vorlage der HSLU	9h
19.03.2025	Umstellung nach der LaTeX-Vorlage der HSLU, Vertiefung in Testing und Erstellung eines Testkonzeptes	5h
20.03.2025	Evaluierung von Anjana & pyCANON, Erstellung weiterer PoCs	4h
23.03.2025	Recherche zu CRF, NER, NLP, Bidirectional Encoder Representations from Transformers (BERT)	4h
24.03.2025	Erweiterung Webportal (Bereich Anonymisierungsmodelle mit Anjana und pyCANON)	2h
25.03.2025	Nachführen der Dokumentation, Einführen einer Teststrategie für den Projektverlauf	3h
29.03.2025	Einbindung von Anjana und pyCANON sowie Überarbeitung der Verarbeitungspipeline (sensitivity_score), README Update, Aktualisieren der Tests	6h
30.03.2025	Recherche Presidio (basierend auf spaCy), Berechnung und Darstellung von Auswertungen der Erkennung und Anonymisierung	6h
31.03.2025	Umstrukturierung des Webportal mit Streamlit Pages, Einbindung von Presidio für unstrukturiert Daten, Unterstützung von unstrukturierten Daten	8h
01.04.2025	Erkennung & Klassifikation von sensiblen Daten in unstrukturiertem Text, Markierung der Textstellen	7h
06.04.2025	Aufbereiten der Arbeiten, Vorbereitung Zwischenpräsentation	4h
07.04.2025	Vorbereitung Zwischenpräsentation	2h
09.04.2025	Vorbereitung, Durchführung und Protokollierung der Zwischenpräsentation	4h
12.04.2025	Nachführung der Dokumentation	3h
13.04.2025	Weiterentwicklung Webportal, Erkennung unstrukturierter Daten, Visualisierung	8h
14.04.2025	Nachführen der Dokumentation	1h

Datum	**Aufgabe**	**Aufwand**
15.04.2025	Implementierung weiterer Anjana-Anonymisierungsmodelle, dynamische Parameter	6h
16.04.2025	Implementierung der Anonymisierungsmodelle und dynamischen Parameter	4h
17.04.2025	Implementierung der Anonymisierungsmodelle und dynamischen Parameter	8h
19.04.2025	Implementierung der Anonymisierungsmodelle und dynamischen Parameter	2h
21.04.2025	Nachführung der Dokumentation	3h
22.04.2025	Verfeinerung der automatischen Erkennung	4h
23.04.2025	Verfeinerung der automatischen Erkennung, Nachführung der Dokumentation	6h
26.04.2025	Überarbeitung der Erkennungs- und Anonymisierungspipeline	8h
27.04.2025	Testung und Evaluierung der Pipeline	1h
28.04.2025	Definition der Art der Ergebnisführung, Nachführen der Dokumentation	5h
29.04.2025	Erweiterung der Dokumentation mit Ergebnissen	4h
30.04.2025	Ausarbeitung der schönen Darstellung von Ergebnissen in der Dokumentation	4h
03.05.2025	Erweiterung der Dokumentation mit Ergebnissen	4h
04.05.2025	Datenverarbeitungspipeline (strukturierte Daten) verschönern	6h
05.05.2025	Datenverarbeitungspipeline (strukturierte Daten) verschönern	8h
06.05.2025	Datenverarbeitungspipeline (strukturierte Daten) verschönern	8h
07.05.2025	Überarbeiten der System-Tests in Python	9h
10.05.2025	Erkennung von unstrukturierten Daten überarbeiten	4h
11.05.2025	Erkennung von unstrukturierten Daten überarbeiten	4h
12.05.2025	Erkennung und Anonymisierung von unstrukturierten Daten überarbeiten	5h
13.05.2025	Testung der Anonymisierungsmethoden (strukturiert), Vergleich der Metriken, Anpassung/Optimierung der Visualisierungen	3h
14.05.2025	Testung der Anonymisierungsmethoden (strukturiert), Vergleich der Metriken, Anpassung/Optimierung der Visualisierungen	5h
17.05.2025	Testung der Anonymisierungsmethoden (strukturiert), Vergleich der Metriken, Anpassung/Optimierung der Visualisierungen, Erstellung Web Abstract Version 1	4h
18.05.2025	Erstellung Web Abstract Version 1	2h
20.05.2025	Besprechung / Sync mit Herr Eskhita, Finalisierung Web Abstract Version 1 für Kontrolle	2h
21.05.2025	Fokus auf Dokumentation	5h
25.05.2025	Fokus auf Dokumentation	8h
26.05.2025	Fokus auf Dokumentation	8h
27.05.2025	Fokus auf Dokumentation	6h
28.05.2025	Fokus auf Dokumentation	6h

Datum	**Aufgabe**	**Aufwand**
29.05.2025	Fokus auf Dokumentation	4h
31.05.2025	Arbeit an Pitch-Video	4h
01.06.2025	Arbeit an Pitch-Video, Finalisierung Dokumentation	7h
02.06.2025	Finalisierung Dokumentation	10h
03.06.2025	Finalisierung Dokumentation, Abgabe der Arbeit	6h
04.06.2025	Finalisierung Dokumentation, Abgabe der Arbeit	4h
Total		**363h**

Tabelle F.1: Arbeitsjournal

G Protokolle

Kick-off (19.02.2025, 1:1 Eskhita / Derungs)

Besprochene Themen	- Vorstellung Projektidee Pascal - Themenbereiche der Arbeit - Grobe Festlegung des Projektscopes - Offene Fragen geklärt
Geplante Arbeiten	- Aufgabenstellung überarbeiten / spezifizieren - Literaturrecherche - Anforderungsanalyse - Use-Case-Definition
Ergebnisse / Erkenntnisse:	- Pascal schickt regelmässige Update-Mails - Wenn gewünscht Meetings vereinbaren, keine fixen Termine - Zitierung mit IEEE in Ordnung - Verwendung von GitHub in Ordnung
Notizen: -	

Tabelle G.1: Protokoll Kick-off 19.02.2025

Nutzwertanalyse / Update (05.03.2025, 1:1 Eskhita / Derungs)

Besprochene Themen	- Besprechung Paarweiser Vergleich Backend/Frontend - Besprechung Nutzwertanalyse Backend/Frontend - Use Cases (Braucht es eine genaue Use Case Beschreibung?) - Anforderungsanalyse / Projektplan (Meilensteine / Issues auf GitHub) - Risikoanalyse
Geplante Arbeiten	- Entscheidungsfindung Programmiersprache / Frameworks - Recherche NLP-Methoden (NER usw...) - Kriterien der Nutzwertanalyse genauer definieren und begründen
Ergebnisse / Erkenntnisse:	- Entscheidungsfindung mit Nutzerwertanalyse nicht ausreichend. Kriterien müssen bewertet werden. - Ist dies die beste Methode? Warum?
Notizen: -	

Tabelle G.2: Protokoll Nutzwertanalyse / Update 05.03.2025

Zwischenpräsentation BAA (09.04.2025, Eskhita / Portmann / Derungs)

Besprochene Themen	- Kurze Vorstellungsrunde - Durchführung der Zwischenpräsentation - Vermittlung der Projektidee an den Experten
Geplante Arbeiten	- Vollendung der Arbeit
Ergebnisse / Erkenntnisse:	- Die Projektidee konnte verständlich übermittelt werden.
Notizen: Bei Fragen technischer Natur, darf der Studierende gerne auch Herr Portmann kontaktieren.	

Tabelle G.3: Protokoll Zwischenpräsentation 09.04.2025

Besprechung/Demo Prototyp (20.05.2025, Eskhita / Derungs)

Besprochene Themen	- Demonstration des Prototyps - Weitere Schritte
Geplante Arbeiten	- Abschluss des Webportals - Fokus auf Dokumentation
Ergebnisse / Erkenntnisse:	- Arbeit ist ön track". - Etwas dem Zeitplan voraus.
Notizen: -	

Tabelle G.4: Protokoll Besprechung/Demo 20.05.2025

Abschlusspräsentation BAA (30.06.2025, Eskhita / Portmann / Derungs)

Besprochene Themen	- Zeigen des Pitch-Videos - Durchführung der Abschlusspräsentation - Verteidigung durch den Studierenden - Feedback / Bewertung an den Studierenden
Geplante Arbeiten	- Keine (Arbeit abgeschlossen).
Ergebnisse / Erkenntnisse:	- Keine (Arbeit abgeschlossen).
Notizen: -	

Tabelle G.5: Protokoll Abschlusspräsentation 30.06.2025