NERMAN

Named-Entity-Recognition Modelle zur Anonymisierung von deutschen Texten

Programm / Ausschreibung	KIRAS, Kooperative F&E-Projekte, KIRAS-Kybernet-Pass CS Kooperative F&E Projekte (CS KFE_2024)	**Status**	laufend
Projektstart	01.10.2025	**Projektende**	31.03.2027
Zeitraum	2025 - 2027	**Projektlaufzeit**	18 Monate
Keywords	Personenbezogene Daten; Textdaten; Named-Entity-Recognition; Anonymisierung		

Projektbeschreibung

Mit der Einführung der Europäischen Datenschutz-Grundverordnung (DSGVO) im Jahr 2018 haben personenbezogene Daten dramatisch an Aufmerksamkeit gewonnen und der Umgang mit diesen Daten ist genau zu hinterfragen. Anonymisierte Daten hingegen sind von der DSGVO ausgenommen, da sie keine Rückschlüsse auf natürliche Personen zulassen.
Das Interesse an Datenanonymisierung ist deshalb stark gestiegen und führte zur Entwicklung verschiedenster Anonymisierungstechniken. Besonders beim Einsatz von KI, wie Prompting für Chatbots oder Training von Large Language Modellen (LLM), ist die Anonymisierung personenbezogener Daten gefragt. Das erfordert geeignete Modelle um ein zuverlässiges und datenschutzkonformes Ergebnis zu gewährleisten. Während für die englische Sprache bereits sehr gute Modelle existieren, ist deren Performance oft mangelhaft, wenn sie auf deutschsprachige Texte angewendet werden.
Übergeordnetes Ziel des Projekts NERMAN ist die Erforschung von Modellen zur
• Identifikation von personenbezogenen Informationen in deutschsprachigen Texten und darauf aufbauend
• Methoden für eine angemessene Anonymisierung der identifizierten Inhalte.
Kernaufgabe ist deshalb die Erforschung von Named-Entity-Recognition (NER) Modellen zur Detektion personenbezogener Inhalte. Dies soll anhand von zwei im Projekt zu definierenden Use Cases umgesetzt werden. Im Speziellen ist die Entwicklung eines NER-Modells geplant, das die Anonymisierung von Texten des BMI ermöglicht, wobei der Fokus auf E-Mail- und Chat-Korrespondenz liegen soll.
Wesentliche Voraussetzung für die Modellentwicklung ist die Gewinnung geeigneter Trainings- und Testdaten. Dabei sollen echte Beispieldaten mit Web-Scraping von öffentlichen Informationen und synthetischer Datengenerierung kombiniert werden. Diese Daten müssen hinsichtlich ihrer Repräsentativität und Eignung bewertet werden. Das soll mittels statistisch-linguistischer Kennzahlen erfolgen. Da aktuell keine zufriedenstellenden deutschsprachigen Datensätze verfügbar sind, ist die Generierung eines deutschsprachigen Benchmark-Datensatzes für ein möglichst breites Spektrum an Anwendungsfällen vorgesehen.
Die entwickelten Modelle werden umfassend validiert und bewertet. Die Bewertung umfasst neben technischen Kriterien wie Performance, Effizienz oder Ressourceneinsatz, auch rechtliche und ethische Faktoren. Das rechtliche und ethische Framework für personenbezogene Daten und Anonymisierungstechniken beim Einsatz von KI soll Metriken zur Bewertung der Qualität einer Anonymisierung beinhalten.

Als Proof-of-Concept werden die besten Modelle in einen zu entwickelnden Demonstrator integriert.
Als wesentliche Innovation des Projekts NERMAN wird erstmals ein NER-Modell entwickelt, das speziell für die Anwendung auf überwiegend deutschsprachige Chat- und E-Mail-Daten zugeschnitten ist.
Eine weitere Neuheit von NERMAN ist die Erstellung von Datensätzen mit ähnlichen linguistischen Eigenschaften wie Chats und E-Mails und dabei speziell die Anwendung von LLMs für die Generierung synthetischer Daten. Erstmals sollen repräsentative, synthetische Testdatensätze, die komplett datenschutzkonform sind, für einen hochsensiblen Sektor wie die Sicherheitsverwaltung generiert und bereitgestellt werden.
Schließlich sollen erstmals quantitative Kriterien erarbeitet werden, die eine möglichst zuverlässige Prüfung des Personenbezugs von Daten und der Qualität von Anonymisierungsvorgängen ermöglichen.

Abstract

The introduction of the European General Data Protection Regulation (GDPR) in 2018 had far-reaching effects on the handling and use of personal data. Anonymized data is exempt from the GDPR, as - ideally - no conclusions can be drawn about natural persons.
In response, global interest in data anonymization has greatly increased, which reflected in the development of various new anonymization techniques. Especially concerning Large Language Models (LLMs), anonymization is of special interest, since it was shown that training data can be extracted retrospectively. To achieve results in accordance with the GDPR, well performing anonymization models are necessary. While many such models exist for the English language, models for German texts are lacking.
The main goal of the NERMAN project is research of machine learning models concerning the
• identification of personalized information in German texts and
• methods for adequate anonymization of the identified data.
To achieve our goal, we plan to develop a Named-Entity-Recognition NER model with a focus on the detection of personalized data. This is to be realised on the basis of two use cases to be defined in the project. A special focus is the anonymization of texts provided by the BMI, which mostly consist of email and chat correspondence.
To develop a performant model, the quality of training data is crucial. To acquire such data, we plan to implement a combination of web-scraping and synthetic data generation. The data will also be compared via statistical metrics to a ground truth to ensure that the data is valid. As there is currently a lack of German NER datasets, we will provide a benchmark dataset based on our acquired data.
The developed models will be thoroughly evaluated in terms of performance, efficiency, resource use and ethical and legal aspects. The ethical and legal framework that will be constructed as a result is also of value for future evaluations of anonymization quality concerning Artificial Intelligence (AI).
In order to demonstrate the practical application of our research, a proof of concept will be constructed.
As a major innovation of the NERMAN project, a NER model tailored for usage with German language email and chat data will be developed for the first time. Furthermore, the generation of synthetic data exhibiting specific text characteristics for NER model training through the utilisation of LLMs is planned. The generated data will be made available to the public via a benchmark data set, which represents a significant development as it is the first time data has been provided for such a highly confidential sector. Lastly, a quantitative framework concerning the evaluation of quality for personal data and anonymization methods will be constructed.

Projektkoordinator

- JOANNEUM RESEARCH Forschungsgesellschaft mbH

Projektpartner

- Axtesys GmbH
- Bundesministerium für Inneres
- Universität für Weiterbildung Krems