RAIDAR RAPID AI BASED DETECTION OF AGGRESSIVE OR RADICAL CONTENT ON THE WEB

Projekt-Newsletter #1

RAPID ARTIFICIAL INTELLIGENCE BASED DETECTION OF AGGRESSIVE OR RADICAL CONTENT ON THE WEB (RAIDAR)

FFG KIRAS 2020

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY

KIRAS Security Research

Liebe Leserinnen und Leser,

dies ist die erste Ausgabe des Newsletters zum Forschungsprojekt *RAIDAR* (Rapid Artificial Intelligence based Detection of Aggressive or Radical content on the Web).

RAIDAR wird im Rahmen des österreichischen Sicherheitsforschungsprogramms *KIRAS* gefördert, einem nationalen Programm zur Förderung der Sicherheitsforschung in Österreich. Die Programmverantwortung für das KIRAS-Programm liegt beim *Bundesministerium für Landwirtschaft, Regionen und Tourismus (BMLRT)*. Das BMLRT hat die *Österreichische Forschungsförderungsgesellschaft (FFG)* mit dem Programm- und Schirmmanagement für das KIRAS-Programm beauftragt.

Gemeinsam mit dem *AIT (Austrian Institute of Technology GmbH)* als leitende Organisation erforschen *Semantic Web Company GmbH*, *SCENOR*, *Research Institute AG & Co KG* und *LIquA - Linzer Institut für qualitative Analysen* Methoden der Erhebung und Bewertung von demokratiegefährdenden Inhalten in großen Datenbeständen, einschließlich Hass und Anzeichen von Radikalisierung. RAIDAR will in diesem Kontext nicht nur neue Methoden zur Sondierung dieser Inhalte liefern, sondern eine anwender*innenfreundliche und IT-basierte Plattform zur teilautomatisierten und versatilen Analyse von großen Datenbeständen aus unterschiedlichsten Quellen entwickeln. Ziel dabei ist, das System in die Lage zu versetzen, automatisch und mit Hilfe von künstlicher Intelligenz relevante Inhalte zu Hass und Anzeichen von Radikalisierung in diesen Datenbeständen zu identifizieren, die aus strafrechtlicher Sicht relevant sein könnten. Als Bedarfsträger fungiert das *Bundesministerium für Justiz (BMJ)*, das durch die Entwicklung dieses teilautomatisierten Assistenzsystems bei seiner juristischen Arbeit entlastet werden soll. Im Rahmen des Forschungsprojekts wird dabei auch eine Technikfolgenabschätzung zu den ethischen Grenzen und rechtlichen Schranken im Kontext von KI-basierter automatisierter Erfassung von Daten durchgeführt.

Weitere Informationen zu den Hintergründen und Zielsetzungen des Projekts finden Sie auf unserer Homepage unter raidar.at/projekt.

Ein kurzer Rückblick auf die letzten Monate und Wochen: Die Arbeiten an *RAIDAR* starteten Ende Oktober 2021 mit einem Kick-Off-Meeting. Daran anschließend folgten ab Dezember 2021 mehrere Workshops, um die Anforderungen des Bedarfsträgers näher zu spezifizieren, die technischen Möglichkeiten – auf Basis der *PoolParty Semantic Suite* - zu erörtern und einen Katalog an Problemen, Risiken und Lösungsansätzen zu erarbeiten. So wurden u. a. Fragen zu Arbeitsabläufen, Datenspeicherung, Leistungsspektrum, Datenexport oder Update-Intervallen besprochen und geklärt. Derzeit wird in diesem Zusammenhang an "User stories"

gearbeitet, um die jeweiligen antizipierten Anwendungsfälle genauer zu beschreiben und eine erste Einschätzung von ethischen und rechtlichen Herausforderungen zu ermöglichen (Stichworte: "ethics by design" und "legal by design").

Sowohl auf technischer, rechtlicher als auch geistes-, sozial- und kulturwissenschaftlicher Seite wurden notwendige Vorarbeiten geleistet, um die Analyse eines brauchbaren Datenbestands zu Hass und radikalisierenden Inhalten in den folgenden Monaten vornehmen zu können. Unter anderem wurden:

- geeignete Content Scraper zum Extrahieren von großen Datenbeständen aus Social-Media-Kanälen entwickelt,
- Experimente zu Text-Corpora-Analyse-Pipelines durchgeführt (Erkennung von Themen und signifikanten Schlüsselwörtern in großen Datenbeständen, Einspeisung in Analyse-Graphen, Modellierung von Beziehungen zwischen Knoten),
- erste Prototypen für das Analyse-Dashboard erstellt (basierend auf Streamlit),
- Erkenntnisse aus aktuellen Forschungen gesichtet, die sich mit der Bereinigung von großen Datenbeständen für die weitere Verarbeitung in Machine-Learning-Modellen auseinandersetzen (z. B. *Davidson et al. 2017*, *Vogel, Regev & Steinebach 2019*, ...),
- diverse Tools für die Annotation von Datenbeständen und den Aufbau von Taxonomien gesichtet (*doccano*, *Prodigy*, *Universal Data Tool*, *Orange*, *Labelbox*, *Humanloop Programmatic*, ...),
- Recherchen zu bestehenden Wortlisten und Taxonomien (*Hatebase*, *Hurtlex*, *Hate Ontology*, HateXplain, ...) sowie annotierten Datenbeständen aus dem Bereich Hassrede und Radikalisierung (*Detecting Offensive Statements Towards Foreigners in Social Media (2017)*, *Overview of the HASOC track at FIRE (2019)*, *RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets (2021)*, …) vorgenommen, um sie für eigene Taxonomien zu adaptieren.

Der RAIDAR-Newsletter wird herausgegeben von:

Austrian Institute of Technology
Giefinggasse 4
1210 Wien
Österreich
Tel.: +43 50 550 - 0
E-Mail: office@ait.ac.at
Website: raidar.at
Autorisierter RAIDAR-Vertreter: Alexander Schindler