RAIDAR

Projekt-Newsletter #2

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

FFG KIRAS 2020

Liebe Leserinnen und Leser,

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

In den letzten beiden Monaten wurde u. a. die Spezifikation der Anforderungen des Bedarfsträgers (Bundesministerium für Justiz) um „User Stories“ erweitert. Damit wurde es möglich, die antizipierten Anwendungsfälle des RAIDAR-Systems genauer zu beschreiben und eine erste Einschätzung von ethischen und rechtlichen Herausforderungen durchzuführen. Im Zuge der rechtlichen Begleitforschung erfolgt derzeit eine eingehende Analyse der Rahmenbedingungen (u. a. unter Behandlung von Aspekten wie anwendbarem Recht, Grundrechtsrisiken im Rahmen der DSFA und Verhältnismäßigkeitsprüfung).

Für die folgende KI-basierte Analyse von Datenbeständen wurden Taxonomien für den Bereich Extremismus sowie für den Bereich Hate Speech fertig gestellt, die auf einer Analyse und Weiterentwicklung bestehender Taxonomien aus diesen Bereichen basieren. Ebenfalls erstellt wurde eine Liste mit über 14.000 Wörtern, die bei Beleidigungen, Beschimpfungen und Hassäußerungen verwendet werden. Sowohl die Wortliste als auch die Taxonomien wurden in die PoolParty Semantic Suite importiert und dort in den folgenden Monaten weiterentwickelt.

Auf technischer Seite wurde mittlerweile ein Content-Scraper implementiert, mit dem Nachrichten aus 20 Telegram-Kanälen mit potenziell extremistischen Inhalten heruntergeladen wurden, die im nächsten Schritt einer KI-basierten Analyse unterzogen werden. Weitere technische Aktivitäten der letzten Wochen umfassen die Bereitstellung und Anwendung von Wissens- und Analysegraphen, die Entwicklung erster auf Machine Learning basierender Module (Named Entity Recognition, Keyword Extraction, Image Similarity), die Implementierung von ersten Dashboards für die Anwender*innen (Keyword Search, Exploratory Analysis Graphs, Image Similarity Search), die Vorbereitung von geeigneten automatischen Detektionsmodellen für Radikalisierung und Hate Speech sowie die Erstellung eines Prototyps für die Erkennung von radikalisierenden Symbolen in Bildinhalten.

Über RAIDAR:

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 - Verein zur Erforschung aktueller gesellschaftlicher Herausforderungen, 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.

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