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

RAIDAR

Projekt-Newsletter #4

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

Die technologische Entwicklung des RAIDAR-Systems wurde in den letzten drei Monaten weiter vorangetrieben. Im Kontext von Wissensextraktion und -management wurde sowohl am Frontend als auch am Backend gearbeitet. Das Frontend wird u. a. eine übersichtliche Strukturierung der analysierten Daten, verschiedene Formen der Visualisierung von Informationen sowie Zusammenfassungen und Empfehlungen bieten, wobei großer Wert auf Usability-Aspekte für die potenziellen Nutzer*innen gelegt wird. Für das Backend wurden u. a. über die Open-Source-Workflow-Management-Plattform „Airflow“ Workflow-Tasks für Sexism-Binary-Detektoren und Toxicity-Binary-Detektoren implementiert. Die bereits trainierten Machine-Learning-Modelle wurden als abrufbare Module in der RAIDAR-Plattform vorbereitet, damit sie gezielt weiter trainiert und angepasst werden können. Weitere Modelle (Extremism-Binary, Threat, Criminal Relevance, Hate-Speech-Binary, ...) sind hier in Arbeit, wodurch am Ende insgesamt rund 20 Detektoren bereitstehen werden.

Seit November 2022 finden alle zwei Monate Workshops mit den Entwickler*innen und potenziellen Nutzer*innen statt, um die technologische Umsetzung aus rechtlicher und ethischer Perspektive zu begleiten. In diesen Workshops wird entlang der zu Projektbeginn definierten User-Stories die rechtliche und ethische Risikoeinschätzung im gesamten Projektteam reflektiert. Dabei erfolgt zu Beginn jeweils eine Vorbemerkung zur Risikoperspektive der jeweiligen User-Story aus Expert*innensicht, um darauf aufbauend über die Risikobeschreibung und -behandlung zu diskutieren. Beispielsweise ergibt sich für die User Story „Aufbau der Taxonomie“ aus der dann geführten Diskussion, dass Taxonomien, die im RAIDAR-System verwendet werden, immer verpflichtend in einem Open-Source-Repository hochgeladen werden müssen. So können u. a. Transparenz, Qualität (Wisdom of Crowd) und Partizipation sichergestellt werden.

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:

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