RAIDAR

Rapid Artificial Intelligence based Detection of Aggressive or Radical content on the Web

Programm / Ausschreibung	KIRAS, Kooperative F&E-Projekte, KIRAS Kooperative F&E-Projekte 2020	**Status**	abgeschlossen
Projektstart	01.10.2021	**Projektende**	30.09.2023
Zeitraum	2021 - 2023	**Projektlaufzeit**	24 Monate
Keywords	Künstliche Intelligenz, Radikalisierung, Hass im Netz		

Projektbeschreibung

Die jüngsten Ereignisse im Zusammenhang mit der US-amerikanischen Präsidentschaftswahl führen uns klar vor Augen, wie virtuell geführte Hasskampagnen reale, demokratie-gefährdende Ereignisse nach sich ziehen können. Geschwindigkeit und Ausmaß dieser Eskalati-on zeigen uns deutlich, dass „Hass im Netz“ und „Radikalisierung“ nicht nur virtuell latente Bedrohungen demokratischer Einrichtungen und der Demokratie selbst darstellen, sondern dass diese durch direkte Angriffe auf Institutionen und Individuen realisiert werden. Ähnliche Phä-nomene und Entwicklungen, wie die gezielte Vermischung radikaler Gruppierungen mit an-fänglich gemäßigten Protestbewegungen, werden auch in Europa und Österreich beobachtet. Diese Vermischung findet nicht nur im öffentlichen, sondern auch im virtuellen Raum statt. Neue digitale Plattformen werden in diesem Kontext in zunehmendem Maße zur Verbreitung demokratiegefährdender Meinungen missbraucht und Hassreden und Hassverbrechen werden zu einem akuten Problem. Eine zentrale Problemstellung im Bereich Hass im Netz stellen die fehlenden Hilfsmittel dar, das Ausmaß dieses Phänomens messbar zu machen. Konkret können somit keine Rückschlüsse über die Reichweite bestimmter Hass-Kampagnen, sowie keine Einschätzung des Problems Hass im Netz im generellen, getroffen werden. In weiterer Folge besteht auch kein Überblick, ob Gegenmaßnahmen zu Hass-Kampagnen Wirkung zeigen.
Ziel des Projekts RAIDAR (Rapid Artificial Intelligence based Detection of Aggressive or Radical content on the Web) ist die Erforschung von Methoden und Ansätzen zur quantitativen Erhe-bung und Bewertung der demokratiegefährdenden Inhalte Hass im Netz und Radikalisierung. Weitere Ziele sind die Entwicklung einer datenwissenschaftlichen Plattform zur teilautomati-sierten und versatilen Analyse großer Datenbestände aus unterschiedlichen Quellen, sowie die Erforschung von Ansätzen und Methoden zur automatisierten Einordnung von Inhalten bzgl. Paragraphen, welche aus strafrechtlicher Sicht Hass im Netz und Radikalisierung zuzuordnen sind. Zu möglichen demokratiegefährdenden Delikten zählen beispielsweise Verhetzung (§ 283 StGB), Wiederbetätigung (VerbotsG) oder Gefährliche Drohung (§ 107 StGB). Wegen der recht-lichen, gesellschaftlichen und kulturellen Komplexität dieser Aufgaben- und Zielsetzung, wird RAIDAR von drei GSK Partnern begleitet, mit dem Ziel einer Umfassenden ethischen und recht-lichen Evaluierung.

Die Innovation von RAIDAR besteht in der Entwicklung und Definition von Kennzahlen, Mess-größen und Methoden zur quantitativen, sowie qualitativen Evaluierung von Hass im Netz und Radikalisierung. Als spezifische Innovation wird die

Anwendung des Forschungsfeldes LegalAI auf den Anwendungsbereich Hass im Netz betrachtet. Im Gegensatz zum Problemfeld Desin-formation im Netz, gibt es klare Paragraphen-Definitionen für die Teilbereiche von Hass im Netz. Diese gut judizierten Paragraphen stellen eine grundlegende Wissensbasis dar, welche durch Kombination von wissensbasierten Systemen mit daten-getriebenen Systemen, in komplexen Analyse- und Erkennungsmethoden abgebildet werden. Projektergebnisse ermöglichten eine Entlastung des Bedarfsträgers (BMJ) durch teilautomati-sierte, auf Künstlicher Intelligenz basierender, Assistenzsysteme im juristischen Bereich. Eine konkrete Technikfolgenabschätzung ethischer Grenzen und rechtliche Rahmenbedingungen im Kontext von Künstlicher Intelligenz zur automatisierten Erfassung von Daten. Anwendung der RAIDAR Plattform in einer quantitativen Studie im Beriech „Hass im Netz“ und „Radikalisierung“ auf zeitlich und kontextuell relevanten Inhalten.

Abstract

The recent events surrounding the U.S. presidential election clearly show us how virtual hate campaigns can lead to real events that threaten democracy. The speed and scale of this escala-tion clearly show us that "online hate" and "radicalization" are not merely virtual latent threats to democratic institutions and democracy itself but are realized through direct attacks on institu-tions and individuals. Similar phenomena and developments, such as the targeted mixing of radical groups with initially moderate protest movements, are also observed in Europe and Aus-tria. This mixing is taking place not only in public space, but also in virtual space. In this context, new digital platforms are increasingly being misused to disseminate opinions that threaten de-mocracy and hate speech and hate crimes are becoming an acute problem. A central problem in the area of Hate Speech is the lack of tools to make the extent of this phenomenon measurable. In concrete terms, it is therefore not possible to draw conclusions about the reach of specific hate campaigns or to assess the problem of Hate Speech in general. As a result, there is also no overview of whether countermeasures to hate campaigns are effective. The main goal of the RAIDAR (Rapid Artificial Intelligence based Detection of Aggressive or Radi-cal content on the Web) project is to research methods and approaches for the quantitative collection and evaluation of hate content and radicalization on the web, both which pose a threat to democracy. Further goals are the development of a data science platform for the semi-automated and versatile analysis of large data sets from different sources, as well as the research of approaches and methods for the automated classification of content according to legal paragraphs, which are to be assigned to hate on the web and radicalization from a criminal law perspective. Possible crimes endangering democracy include incitement to hatred (§ 283 StGB), re-invasion (VerbotsG) or dangerous threats (§ 107 StGB). Due to the legal, social and cultural complexity of these tasks and objectives, RAIDAR research is accompanied by three GSK partners with the aim of a comprehensive ethical and legal evaluation.
The innovations of RAIDAR include the development and definition of metrics, measures and methods for quantitative and qualitative evaluation of online hate and radicalization. The ap-plication of the research field LegalAI to the application area of Hate Speech is considered a specific innovation. In contrast to the problem area of disinformation on the internet, there are clear legal definitions for the sub-areas of Hate Speech. These well-judged legal paragraphs rep-resent a fundamental knowledge base that will be mapped into complex analysis and detection methods by combining knowledge-based systems with data-driven systems.
Project results will include a reduction of the workload of the BMJ (Federal Ministry of Justice) by means of partially automated assistance systems based on artificial intelligence in the legal field, a concrete technology assessment of ethical limits and legal frameworks in the context of artificial intelligence for automated data collection and the application of the RAIDAR platform in a quantitative study in the area of "Hate Speech" and "radicalization" on temporally and contex-tually relevant content.

Projektkoordinator

- AIT Austrian Institute of Technology GmbH

Projektpartner

- Research Institute AG & Co KG
- Bundesministerium für Justiz
- Scenor - Verein zur Erforschung aktueller gesellschaftlicher Herausforderungen
- LIquA - Linzer Institut für qualitative Analysen
- Semantic Web Company GmbH