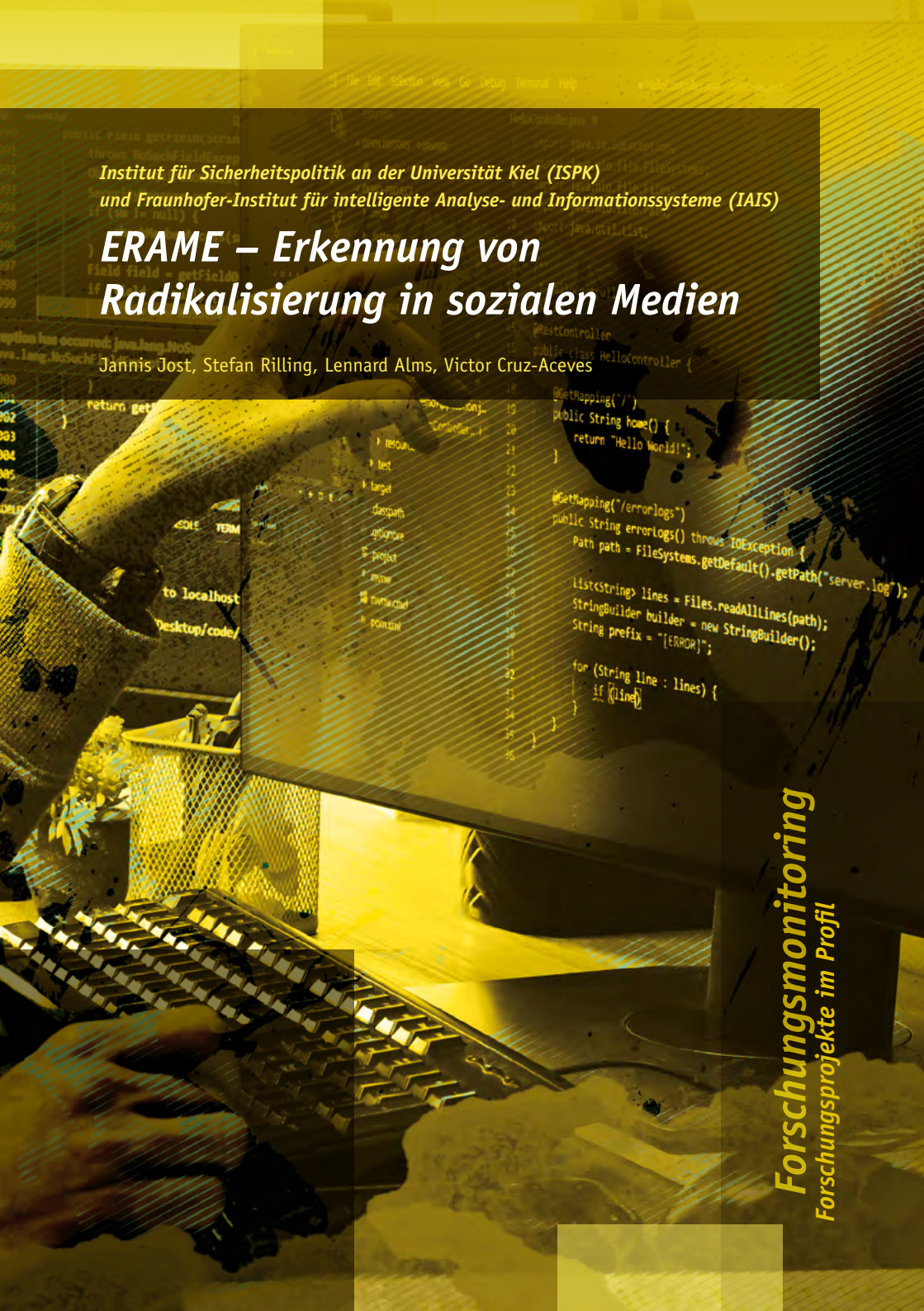
Institut für Sicherheitspolitik an der Universität Kiel (ISPK)
und Fraunhofer-Institut für intelligente Analyse- und Informationssysteme (IAIS)

ERAME – Erkennung von Radikalisierung in sozialen Medien

Jannis Jost, Stefan Rilling, Lennard Alms, Victor Cruz-Aceves



Forschungsmonitoring
Forschungsprojekte im Profil

Einleitung, Problematik, Anwendungsfall

Hinterher erscheint immer alles ganz offensichtlich: Im Rahmen der Israel-feindlichen Proteste nach dem 7. Oktober 2023, bei denen es in Deutschland zu Drohungen, Volksverhetzung und zu gewaltsamen Ausschreitungen kam, wurde der Eskalation in entsprechenden Social-Media-Kanälen mittels Schuldzuweisungen und einseitiger Darstellungen der Boden bereitet (Dittrich et al., 2024). Vor den von Rechtsextremist*innen durchsetzen Anti-Corona-Demonstrationen auf den Stufen des Reichstags am 29. August 2020 kursierten etliche diesbezügliche Pläne, Aufrufe und Autopropaganda-Inhalte auf den einschlägigen Online-Plattformen und -Kanälen (Petersen & Groeneveld, 2020). Und auch in der Online-Historie etlicher Terrorist*innen oder Gewalttäter*innen findet sich Hassrede gegen die Menschengruppen, die sie später physisch angriffen – wie im Falle des Terroranschlags von Christchurch im Jahr 2019 (Veilleux-Lepage et al., 2020) oder des Amoklaufs in Plymouth im Jahr 2021 (White, 2021). Die Sicherheitsbehörden stehen vor der Problematik, den Überblick zu behalten, wie sich die verschiedenen extremistischen Phänomenbereiche online entwickeln, welche Themen und Trends eine besondere Dynamik entfalten und, idealerweise, wo diese Dynamik in realen Taten zu gipfeln droht.

Angesichts der Unmenge an Beiträgen, die in den sozialen Medien veröffentlicht werden, ist es mit einem vertretbaren Personalaufwand kaum noch möglich, einen entsprechenden Überblick zu wahren. Seit mehr als drei Jahren forschen die BMBF-geförderten Projekte ERAME und ERAME-REX deswegen daran, wie ein bedarfsgerechtes und praxisnahes Tool aussehen könnte, das es den Verfassungsschutzbehörden ermöglicht, mittels Künstlicher Intelligenz (KI) „vor der Lage“ zu bleiben, was radikale Dynamiken auf sozialen Medien angeht. Ziel ist es, durch eine multidimensionale Vorauswertung 1) relevante Trends und Themen schnell zu erkennen und aufzubereiten und 2) die relevantesten Inhalte und Akteure für die anschließende Auswertung durch Fachpersonal zu priorisieren. Dadurch soll den menschlichen Analyst*innen eine bessere Informationsgrundlage zur Priorisierung von Fällen zur Verfügung gestellt werden, die gleichzeitig mehr zeitliche Ressourcen für die Bearbeitung der wichtigsten Entwicklungen und Akteure freisetzt.

Zur Realisierung haben das Institut für Sicherheitspolitik an der Universität Kiel (ISPK) als sozialwissenschaftlicher Partner, das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) als technologiewissenschaftlicher Partner und das Centre for Security und Society an der Universität Freiburg (CSS) als rechtswissenschaftlicher Partner mit dem Landesamt für Verfassungsschutz Baden-Württemberg und dem Landeskriminalamt Niedersachsen als Praxispartner ein Konsortium gebildet, um verschiedene Ansätze zu untersuchen und zu evaluieren. Konkret wurden zwei Demonstratoren entwickelt, von denen einer für islamistische Akteure auf YouTube und der andere für rechtsextreme und neurechte Akteure auf Telegram optimiert ist. Im Folgenden wird ein Einblick in ihre Funktionsweise und Performanz gegeben.

Systemaufbau

Die entwickelten Demonstratoranwendungen sind als verteilte Systeme konzipiert. Die einzelnen Komponenten zur Datenakquise, Datenverwaltung und Datenverarbeitung sind dabei auf von Fraunhofer IAIS zur Verfügung gestellten Servern installiert. Diese stellen auch die grafische Benutzeroberfläche als Webanwendung zur Verfügung. Sämtliche Komponenten können durch die Docker-Virtualisierung auf beliebigen Servern bereitgestellt werden.

Datenakquise und Datenmanagement

Zur Datenakquise wurden ein Crawler für YouTube und für Telegram als containerisierte, konfigurierbare Backend-Anwendung umgesetzt. Diese arbeiten eine Liste von konfigurierten YouTube-Kanälen und Telegram-Gruppen/-Kanäle alle 24 Stunden ab. Bei jedem Durchlauf werden alle neuen Kommentare in allen Videos beziehungsweise alle neuen Telegram-Nachrichten aus der Liste analysiert (Text, Autor*in, Erstelldatum, Likes, Anzahl der Antworten) und im Datenmanagementsystem persistiert. Der Web-Crawler nutzt die offiziellen YouTube- und Telegram-Programmierschnittstellen.

Das Datenmanagementsystem wird mit der dokumentbasierten Datenbank MongoDB umgesetzt. Mittels der Softwarebibliothek RESTHeart werden REST-basierte Schnittstellen zur Datenbank bereitgestellt. Die

Datenbank dient zur Speicherung der Kommentardaten und der Analyseergebnisse, aber auch zur Verwaltung von Nutzerzugängen und zur Konfiguration der Benutzeroberfläche.

KI-Integration

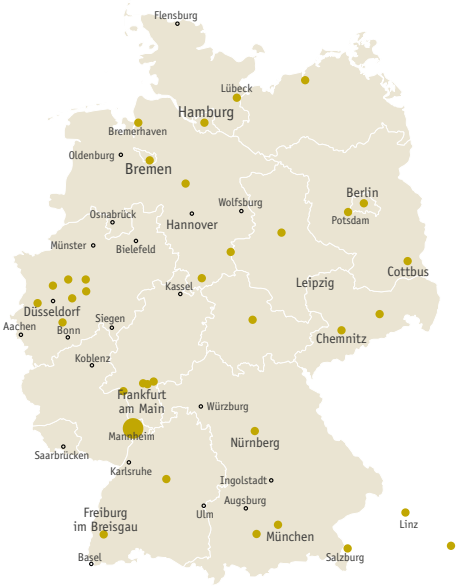
Die verschiedenen KI-Modelle werden nach der Datenakquise automatisch auf allen neuen Kommentartexten ausgewertet. Die Resultate werden mit den Kommentaren in die Datenbank geschrieben.

Grafische Benutzeroberfläche

Die Benutzeroberfläche bietet verschiedenste Sichten auf die Daten und viele Möglichkeiten, diese zu filtern, um detaillierte Erkenntnisse zu gewinnen. Auf verschiedenen Analyseseiten sind die Ergebnisse der KI-Methoden grafisch oder tabellarisch aufbereitet. Dazu gehören unter anderem:

- Die Markierung von diskutierten Orten auf einer Karte mit größeren Punkten für stark diskutierte Orte (Abbildung 1).
- Eine Visualisierung der diskutierten Themen, welche die Häufigkeit und den Zeitverlauf darstellt (Abbildung 2).
- Ein Diagramm, das die Anzahl der neuen Posts pro Woche/Tag und die aggregierten Sentimentverläufe anzeigt.

Des Weiteren existieren Seiten zur Darstellung der Posts/Nachrichten/Kommentare, der Nutzer*innen und der Gruppen/Kanäle mit Möglichkeiten zum Suchen, Filtern und Sortieren.



Geographische Orte

Ort	Häufigkeit
Deutschland	2 881
Österreich	365
Italien	101
Ukraine	93
USA	71
Russland	58
Frankreich	36
Schweiz	29

Abbildung 1: Grafische Darstellung der geografischen Orte, die in einer Telegram-Gruppe in einer Kalenderwoche diskutiert wurden.

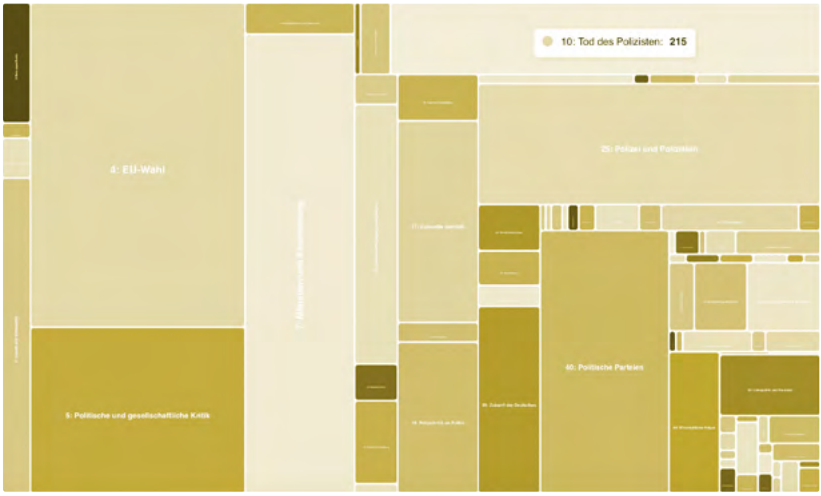


Abbildung 2: Visualisierung der in einer Telegram-Gruppe in einer Kalenderwoche diskutierten Themen.

Operationalisierung sozialwissenschaftlicher Modelle

Die Einstufung, ob ein Inhalt als extremistisch zu bewerten ist oder nicht, ist eine Kernaufgabe der Verfassungsschutzbehörden, die nicht auf eine KI ausgelagert werden, kann oder soll. Um stattdessen menschliche Analyst*innen bestmöglich bei ihren Einstufungen zu unterstützen, wurden zwei Kategoriesysteme entwickelt, die Elemente von politischer Radikalität in unterschiedlichen Ausprägungen (0–5) abbilden. Einige dieser Kategorien sind phänomenbereichsübergreifend („Gewaltaffinität" und „Gruppenbezogene Menschenfeindlichkeit"), andere phänomenbereichsspezifisch (für Rechtsextremismus zum Beispiel „Polarisierung & Elitenfeindlichkeit", „Ablehnung demokratischer Normen & Entscheidungen" sowie „Ablehnung von Meinungs- und Mediendiversität"). Obwohl im Fall des extremistischen Islamismus die gleichen abstrakten Grundwerte berührt sind, wurden spezifische Kategorien formuliert: „Ablehnung demokratischer Normen & Entscheidungen" entspricht beispielsweise sinngemäß der Islamismus-Kategorie „Demokratieverachtung und Gottessouveränität als politisches Prinzip".

Dieses Vorgehen erlaubt es, jede Ausprägung jeder Kategorie mit einer möglichst konkreten, phänomenbereichsspezifischen Definition zu hinterlegen, die mit den Endanwender*innen abgestimmt ist und sich in deren Arbeitsrealität wiederfindet. Die Definition der Ausprägung 3 der Islamismus-Kategorie „Demokratieverachtung und Gottessouveränität als politisches Prinzip" lautet beispielsweise: *„Die Trennung von Politik und Religion ist unislamisch, deswegen ist Distanz zur deutschen Politik angebracht; die deutsche Politik ist Muslim*innen gegenüber grundsätzlich feindlich eingestellt."* Diese mit kurzen Definitionen hinterlegten Kategoriesysteme sind hilfreich, um a) die Einstufung des Expertenteams bei der händischen Annotation (bestmöglich) zu vereinheitlichen und b) bei Prompt-Engineering den Suchauftrag pointiert zu formulieren. Für die Endanwender*innen ergibt sich c) der Vorteil, dass keine übersimplifizierte, binäre Einstufung „extremistisch/nichtextremistisch" vorgenommen wird, sondern ein Textbeitrag auf einem Spektrum verortet wird, das der Multidimensionalität von Extremismus Rechnung trägt.

Der Auftrag der Verfassungsschutzbehörden endet nicht mit der Einstufung von Inhalten oder Akteuren hinsichtlich etwaigen Extremismus. Auch die

Gefahrenabwehr durch die frühzeitige Identifikation von möglichen Straf- und insbesondere Gewalttäter*innen ist ein wichtiger Teil der Arbeit. Da diese Gefahrenanalyse sozialwissenschaftlich noch erheblich komplexer und rechtlich eingriffsintensiver ist, wurde sie im Rahmen von ERAME und ERAME-REX nicht prioritär verfolgt. Dennoch wurden Grundlagen entwickelt, auf die später aufgebaut werden kann, und auch, um bei den Endanwender*innen das Bewusstsein dafür zu schärfen, dass, in den Worten von McCauley und Moskalenko, „[…] es wichtig ist, zwischen Radikalisierung von Meinung und Radikalisierung von Handlung zu unterscheiden" (McCauley & Moskalenko, 2017, S. 264). Dazu wurden Kategoriesysteme wie die oben beschriebenen entwickelt, hier allerdings mit psychologischen anstatt politischen Merkmalen. Bei der Entwicklung der Kategorien dienten die Indikatoren des Risikobewertungsinstruments TRAP-18 als Vorbild (Meloy et al., 2012), das bereits explorativ auf Social-Media-Inhalte angewendet wurde (Cohen et al., 2014).

KI-Verfahren

Im Rahmen des Vorhabens wurden der Trainingsprozess und eine weitgehend automatisierte Trainingspipeline konzipiert und umgesetzt. Durch diese Maßnahmen konnte das KI-Verfahren effizient trainiert werden. Zum Ende des Vorhabens konnten so circa 2 400 Kommentare vom Expert*innenteam des ISPK annotiert werden, wobei jeder Kommentar von drei Personen bearbeitet wurde.

Zur Klassifikation von Kommentaren oder Nachrichten hinsichtlich ihrer Radikalität wurde ein Transformer-basiertes, multilinguales XLM-RoBERTa-Modell (Conneau et al., 2019) in die Anwendung integriert. Das Modell wurde dann mittels des annotierten Kommentardatensatzes einem sogenannten Feintuning unterzogen. Das Modell wurde darauf trainiert, fünf ausgewählte inhaltliche Radikalitätsindikatoren im Kommentartext zu erkennen.

Ein weiteres Transformer-basiertes Modell basierend auf BERT (Chan et al., 2020) wurde mit einem „Learning to Rank"-Ansatz trainiert. Dabei wurde eine „Normalized Discounted Cumulative Gain"-Fehlerfunktion (Wang et al., 2018) mit dem Ziel optimiert, dass die Texte nach der Radikalität sortiert werden können. Auch hier wurde das Modell trainiert, die fünf ausgewählten inhaltlichen Radikalitätsindikatoren abzubilden.

Des Weiteren wurde eine Programmierschnittstelle zu einem GPT4-Modell verwendet, um durch Prompt-Engineering die Texte auf erwähnte Themen, Personen, Gruppierungen, Landmarken und Orte zu untersuchen. Die einzelnen Themen wurden dann über ein Vector-Embedding in übergeordnete Themencluster durch eine hierarchische Clusteranalyse eingeteilt. Für erwähnte Landmarken wie zum Beispiel das Brandenburger Tor wurden durch die Programmierschnittstelle von OpenStreetMap GPS-Koordinaten, die Stadt und das Land ermittelt und als Metadaten abgespeichert. GPT4 wurde für diesen Anwendungsfall mit lokalen Large Language Models und Embedding Models verglichen. GPT4 war dabei augenscheinlich besser, aber eine komplett lokale Lösung ist generell mit ähnlichen Resultaten möglich.

Evaluation

Generell ist die Aufgabe zu entscheiden, ob ein Text radikal ist oder nicht, keine leichte. Die Analysen der Expert*innenannotationen zeigen, dass diese sich auch häufig uneinig sind. Dies liegt insbesondere daran, dass es keine klare Trennlinie, sondern einen großen Graubereich mit viel Interpretationsspielraum gibt. Dies macht das Training von KI-Systemen zur Erkennung von Radikalität schwer und muss bei der Verwendung berücksichtigt werden. Dazu kommt, dass die Radikalität der Texte in dem verwendeten Datensatz (600 000 YouTube-Kommentare und 1 000 000 Telegram-Nachrichten) exponentialverteilt ist. Das heißt sehr viele nicht radikale Texte, wenige leicht radikale Texte und sehr wenige extremistische Texte. Dies spiegelte sich in mehreren Annotationsrunden der Experten wider und führte zusätzlich zu einem sehr unausgewogenem Trainingsdatensatz für die KI-Systeme. Daraus resultierte eine zu erwartende hohe Spezifität (~98 %) und eine geringe Sensitivität (~50 %). Dies spiegelt in etwa auch die Übereinstimmung der Expert*innen untereinander wider. Daraus kann man schlussfolgern, dass es für einen Großteil der Texte offensichtlich ist, dass diese nicht radikal sind, und, dass die restlichen Texte häufig nicht klar eingeordnet werden können.

Zwei Methoden wurden identifiziert, um die Modelle trotz der beschriebenen Schwierigkeiten effektiv zu operationalisieren: Aggregation und Ranking.

Aggregierte Radikalitätsindikatoren

Durch die hohe Spezifität kann davon ausgegangen werden, dass das Verhältnis von radikalen zu gemäßigten Texten im Durchschnitt annähernd korrekt von der KI ermittelt wird. Berechnet man dieses Verhältnis mithilfe der KI für jeden Tag oder jede Woche, so sieht man ein gewisses Grundrauschen, aber auch diverse Spitzen, welche nicht durch die Modellungenauigkeit erklärt werden können. Diese Spitzen weisen auf polarisierende Ereignisse hin. Abbildung 3 und 4 zeigen die Ergebnisse hierzu. Man sieht insbesondere:

1. Nach dem Angriff der Hamas auf Israel schlagen gruppenbezogene Menschenfeindlichkeit und Gewalt stark aus.

2. Nach der Veröffentlichung des Correctiv-Berichtes „Geheimplan gegen Deutschland“ steigt die Elitenfeindlichkeit im rechtsextremen Lager auf ihren Jahreshöchstwert.

3. Nach den Kundgebungen von Muslim Interaktiv in Hamburg steigt die Demokratieverachtung und Gottessouveränität als politisches Prinzip.

4. Nach dem Mord des Polizisten in Mannheim steigt die Gewaltaffinität auf ihren Jahreshöchstwert und die Elitenfeindlichkeit sinkt gleichzeitig auf das Jahresminimum.

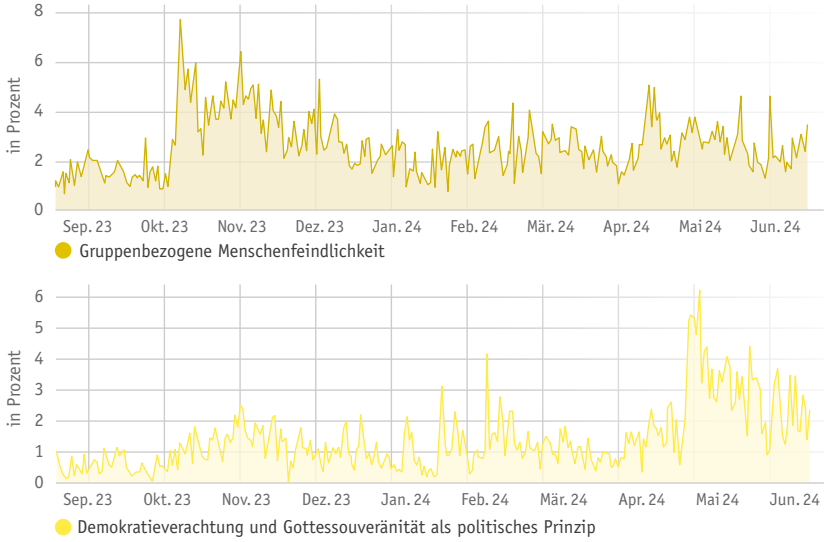


Abbildung 3: Radikalitätsverlauf über die Zeit für ausgewählte Indikatoren im Phänomenbereich PMK Islamismus. Deutlich zu sehen sind die Ausschläge für den 7. Oktober 2023 (oben: Terrorangriff der Hamas auf Israel) und Ende April 2024 (unten: Kundgebung von Muslim Interaktiv in Hamburg). Gemessen wird der Anteil der für den jeweiligen Indikator als „hoch" eingestuften Kommentare.

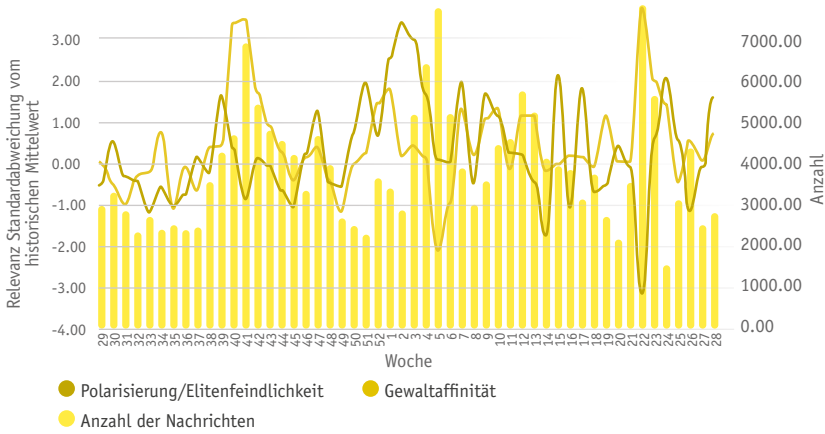


Abbildung 4: Radikalitätsverlauf über die Zeit für ausgewählte Indikatoren im Phänomenbereich PMK Rechtsextremismus. Deutlich zu sehen sind die Ausschläge für den 7. Oktober 2023 (KW 40, Terrorangriff der Hamas auf Israel), den 10. Januar 2024 (KW 2, Correctiv-Bericht „Geheimplan gegen Deutschland") und der 31. Mai 2024 (KW 22, Mord des Polizisten in Mannheim). In Gelb hinterlegt ist die Anzahl der Nachrichten. Gemessen wird die Standardabweichung von dem historischen Durchschnitt der Textrelevanz für den jeweiligen Indikator.

Ranking

Da im Einzelfall der Grad der Radikalität eines Textes immer von einzelnen Interpretationsweisen abhängt, bietet eine Sortierung von Texten eine nuancierte Sicht auf die Relevanz der Texte. Das KI-Verfahren trifft dabei keine binäre Entscheidung, sondern Texte werden im Verhältnis zueinander betrachtet. Dies ermöglicht es nicht nur, die radikalen Texte, sondern die radikalsten Texte zu finden. Zusätzlich suggeriert die Sortierung dem*der Anwender*in keine endgültige Einstufung, womit der*die Anwender*in weniger beeinflusst wird. Durch die Sortierung des gesamten Datensatzes (1 000 000 Telegram-Nachrichten) nach Gewaltaffinität konnten Morddrohungen, Aufrufe zur Gewalt gegen Personengruppen und andere Gewaltfantasien unmittelbar gefunden werden. Eine Schwierigkeit für das KI-System besteht hier insbesondere darin, zu unterscheiden, ob es sich bei dem Text um ein Zitat, einen Bericht oder um die Meinung des*der Autor*in handelt. Dies führt dazu, dass zum Beispiel Berichte über Gewalt im Ukraine-Russland-Krieg sich recht weit oben in der Sortierung für Gewaltaffinität wiederfinden. Diese Berichte sind jedoch unter dem Gesichtspunkt der Radikalität nicht relevant. Ein größerer Trainingsdatensatz, welcher viele dieser Beispiele enthält, könnte in diesem Fall zu einer Verbesserung des Modells beitragen.

Anwendungsbezug

Die vorgestellte Demonstratoranwendung kann durch die aggregierten Radikalitätsindikatoren und die Erkennung von Themen, Personen, Gruppierungen, Landmarken und Orten zur Lagebilderstellung über die extremistischen Szenen einen großen Beitrag liefern. Des Weiteren ist die Sortierung nach einzelnen oder aggregierten inhaltlichen Radikalitätsindikatoren ein signifikanter Mehrwert für die Analyst*innen, welche sich täglich mit der Sichtung dieser Inhalte beschäftigen. Dies konnte in mehreren Anwender*innengesprächen bestätigt werden. Es zeigt sich einmal mehr, dass KI keinesfalls Menschen „ersetzen“ kann in dem Sinn, dass die Qualität mit der menschlicher Analysen vergleichbar wäre. Aber sie kann sehr wohl menschliche Analyseprozesse augmentieren und vorher dazu beitragen, dass menschliche Ressourcen effizienter verwendet werden können.

Literatur

Chan, B., Schweter, S. & Möller, T. (2020): German›s next language model. *arXiv preprint arXiv:2010.10906*

Cohen, K., Johansson, F., Kaati, L. & Clausen Mork, J. (2014): Detecting linguisitic markers for radical violence in social media. *Terrorism and Political Violence*, *26*(1), 246–256. http://dx.doi.org/10.1080/09546553.2014.849948

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2019): Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*

Dittrich, L., Ettmüller, E., Eigenbrodt, S., Hoffmann, F. & Schleinzer, A. (2024, 26. April): *Der 7. Oktober 2023 und seine Folgen.* Gastbeitrag auf Springer Professional. https://www.springer-professional.de/de/verwaltungsmanagement/der-7--oktober-2023-und-seine-folgen/26984032

McCauley, C. & Moskalenko, S. (2017): *Friction. How conflict radicalizes them and us. revised and expanded edition.* Oxford University Press.

Meloy, R. J., Hoffmann, J., Guldimann, A. & James, D. (2012): The role of warning behaviors in threat assessment: an exploration and suggested typology. *Behavioral Science and the Law, 30*(3), 256-279. https://doi.org/10.1002/bsl.999

Petersen, L. & Groeneveld, J. (2020, 20. August): *Geheime Chat-Instruktionen für „Sturm auf den Reichstag": Sicherheitsbehörden waren vor Reichstagsausschreitungen gewarnt.* Business Insider. Abgerufen am 21. Juli 2024 von https://www.businessinsider.de/politik/deutschland/chat-instruktionen-sturm-berlin-verfassungsschutz-sieht-gefahr-durch-corona-demonstranten/

Veilleux-Lepage, Y., Daymon, C. & Amarasingam, A. (2020): *The Christchurch Attack Report: Key Takeaways on Tarrant's Radicalization and Attack Planning.* ICCT Perspective, Dezember 2020. https://www.icct.nl/sites/default/files/2022-12/Christchurch-report-Dec-2020_Spelling-fixed.pdf

Wang, X., Li, C., Golbandi, N., Bendersky, M. & Najork, M. (2018): The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM international conference on information and knowledge management* (S. 1313–1322).

White, M. (2021, 14. August): *Plymouth shooting: Jake Davison liked gun videos and talked about 'incel' in the weeks before attack.* https://news.sky.com/story/plymouth-shooting-man-suspected-of-killing-five-people-and-himself-named-as-jake-davison-23-12380132