Abschlussbericht
Teil II: Eingehende Darstellung

Verbundprojekt:	Einsatz von KI zur Früherkennung von Straftaten
Akronym:	KISTRA
Teilvorhaben:	Teilvorhaben: Technischer Software-Demonstrator
Förderkennzeichen:	13N15345
Laufzeit des Vorhabens:	01.07.2020 bis 31.12.2023
Datum:	28.06.2024
Ausführende Stelle:	Munich Innovation Labs GmbH Pettenkoferstr. 24 80336 München
Ansprechpartner:	Dr. Stefan Taing st@munich-innovation.com 017610426402

1 Einleitung

Gegenstand des vorliegenden Dokumentes ist die ausführliche Darstellung der Inhalte des geförderten Forschungsprojektes 'Einsatz von KI zur Früherkennung von Straftaten (KISTRA)', Teilvorhaben ''Technischer Software-Demonstrator (KISTRA-TSD)'.

Bevor auf die durchgeführten Arbeiten eingegangen wird, werden im Folgenden die Zielsetzung des Forschungsprojektes, die beteiligten Partner und Teilvorhaben des Gesamtvorhabens 'Einsatz von KI zur Früherkennung von Straftaten (KISTRA)' sowie die Kurzfassung der Zielsetzung und Ergebnisse des Teilvorhabens 'Technischer Software-Demonstrator (KISTRA-TSD)' dargestellt.

Zielsetzung des Projekts und beteiligte Partner

Ziel von KISTRA war die Erforschung der Möglichkeiten und Rahmenbedingungen für den ethisch und rechtlich vertretbaren Einsatz von künstlicher Intelligenz (KI) durch polizeiliche EndanwenderInnen zur Erkennung, Vorbeugung und Verfolgung von Straftaten. In KISTRA werden technische Lösungen im Kriminalitätsfeld „Hasskriminalität“ und im Einsatzfeld „Verarbeitung von Bilddaten in Sicherheitsbehörden“ erarbeitet und erprobt.

Die Betrachtung der Rechtmäßigkeit und der ethischen Vertretbarkeit der angestrebten KI-Lösungen und der daraus resultierenden Arbeitsroutinen für die Polizei erfolgt hierbei durch die Johannes-Gutenberg-Universität Mainz (JGU) und die Technische Universität Berlin (TUB). Die sozialwissenschaftliche Betrachtung der Wahrnehmung und des Umgangs mit politisch motivierter Hassrede und Hasskriminalität im Internet erfolgt durch die Ludwig-Maximilians-Universität München (LMU). Die Erarbeitung und Implementierung von adaptiven KI-Methoden zur Verarbeitung von Massendaten für polizeiliche Anwendungen erfolgt durch die Zentrale Stelle für Informationstechnik im Sicherheitsbereich (ZITiS), die Universität Duisburg-Essen (UDE) bzw. FernUniversität Hagen sowie durch die Firma Munich Innovation Labs GmbH (MIL). Die Überprüfung der Robustheit von solchen KI-Modellen gegenüber Angriffen erfolgt durch die Technische Universität Darmstadt (TUD). Ethische Fragen im Zusammenhang mit dem Einsatz von Methoden der künstlichen Intelligenz im polizeilichen Kontext und zum Datenschutz sowie die Durchführung von Nutzertests werden durch die Rheinisch-Westfälische Technische Hochschule Aachen bearbeitet. Die Integration der technischen Teillösungen in einen Gesamtdemonstrator, welcher die besonderen Anforderungen des Betriebes in einer polizeilichen Infrastruktur berücksichtigt, erfolgt schließlich durch MIL. Die polizeilichen Endanwender werden durch das Bundeskriminalamt (BKA) vertreten.

Zielsetzung des Teilvorhabens

Ziel des Teilvorhabens KISTRA-TSD ist die Erforschung und Bereitstellung von technischen Lösungen, die Algorithmen der KI nutzen, um insbesondere Hassrede und verwandte multimediale Inhalte wie Bild- und Videodaten zu klassifizieren und somit eine Priorisierung in der polizeilichen Auswertung zu ermöglichen, um die Effizienz der Bearbeitung im Angesicht der großen und unübersichtlichen Datenmengen zu steigern. Insbesondere ist der rechtliche und ethische Kontext der Datenverarbeitung im Rahmen der Arbeit von

Sicherheitsbehörden zu beachten und - in Zusammenarbeit mit den Projektpartnern - bereits in der vorgeschlagenen technischen Lösung zu adressieren.

Die Entwicklung von immer leistungsfähigeren KI-Algorithmen für die Text- und Bilderkennung wird international von führenden Universitäten und privatwirtschaftlichen Unternehmen getrieben, die vor allem privatwirtschaftliche Anwendungen im Fokus sehen. Dies gilt sowohl für die einzelnen algorithmischen Ansätze als auch für die Infrastrukturkomponenten, die eine effiziente Cloud-basierte verteilte Berechnung ermöglichen. Diese Lösungen stellen sich allerdings für die besonderen Zwecke von Behörden und Organisationen mit Sicherheitsaufgaben (BOS) sowohl hinsichtlich der Anwendungsfälle als auch hinsichtlich rechtlicher Fragestellungen wie z.B. dem Datenschutz als nicht ausreichend heraus. Im Rahmen von KISTRA-TSD werden daher Ansätze zur Textklassifizierung, Ansätze zur Bildklassifizierung, sowie im Kontext der Erstellung des Gesamtdemonstrators Ansätze zur Bereitstellung von Infrastrukturkomponenten für die Zwecke der BOS erforscht. Ziel von KISTRA-TSD ist es hierbei, diese für die existierenden Algorithmen für BOS-spezifische Anwendungen nutzbar zu machen.

2 Durchgeführte Arbeiten

Im Folgenden werden die durchgeführten Forschungsarbeiten im Vergleich zur ursprünglichen Vorhabenbeschreibung ausführlich dargestellt.

Die technischen Lösungen, die im Rahmen von KISTRA-TSD erarbeitet werden, basieren auf dem KI-Framework „MIL.ml Net“ und der OSINT-Analyseplattform „INspectre“, die von der Firma Munich Innovation Labs GmbH (MIL) im Rahmen von Forschungsprojekten zur Auswertung von Daten aus sozialen Netzwerken (Projekte INTEGER und PANDORA) und Kundenprojekten für die Online-Datenverarbeitung von sensitiven Bilddaten, wie beispielsweise medizinischen Bilddaten, erarbeitet wurden und weiterentwickelt werden.

Bedarfsanalyse und Nutzeranforderungen

Die für KISTRA-TSD spezifischen, technischen, ethischen und rechtlichen Anforderungen wurden zunächst im Rahmen einer Bedarfsanalyse innerhalb der Arbeitspakete (AP) 1, 2 und 3 in Form eines Workshops mit den polizeilichen Endanwendern erfasst, priorisiert und dokumentiert. Insgesamt wurden 111 Einzelanforderungen (davon 53 mit hoher Priorität) aufgenommen. Die wichtigsten Anforderungen umfassen neben KI-unterstützten Klassifizierungsmöglichkeiten der verschiedenen Ausprägungen von Hasskriminalität, deren Themenfelder und Ideologien die Integrierbarkeit des Gesamtdemonstrators in existierende polizeiliche Infrastrukturen, die Unterstützung von gängigen Dateiformaten sowie grundlegende Interaktionsmöglichkeiten (Sortierung, Filterung, Suchfunktionen) mit den Daten bzw. den Analyseergebnissen. Weitere wichtige Anforderungen waren das Erkennen der Originalsprache eines Inhaltstextes sowie Erklärungen zu den einzelnen Modellvorhersagen. Es wurde ein technisches Anforderungsdokument erstellt, das zum einen die rechtlichen und ethischen Aspekte des Forschungsvorhabens grundsätzlich umrahmt und zum anderen das Design des technischen Systems hinsichtlich Funktionalitäten, Schnittstellen und Mensch-Technik-Interaktion beschreibt [1].

Abb. 1: Annotationsklassen und Beispiele für den Trainingsdatensatz zur automatischen Erkennung von Hakenkreuzen als politisches Symbol des Nationalsozialismus.

Rechts	Religiös	Antifa	Uneindeutiger Kontext	Unvollständig	Insgesamt
818	640	74	846	495	2873

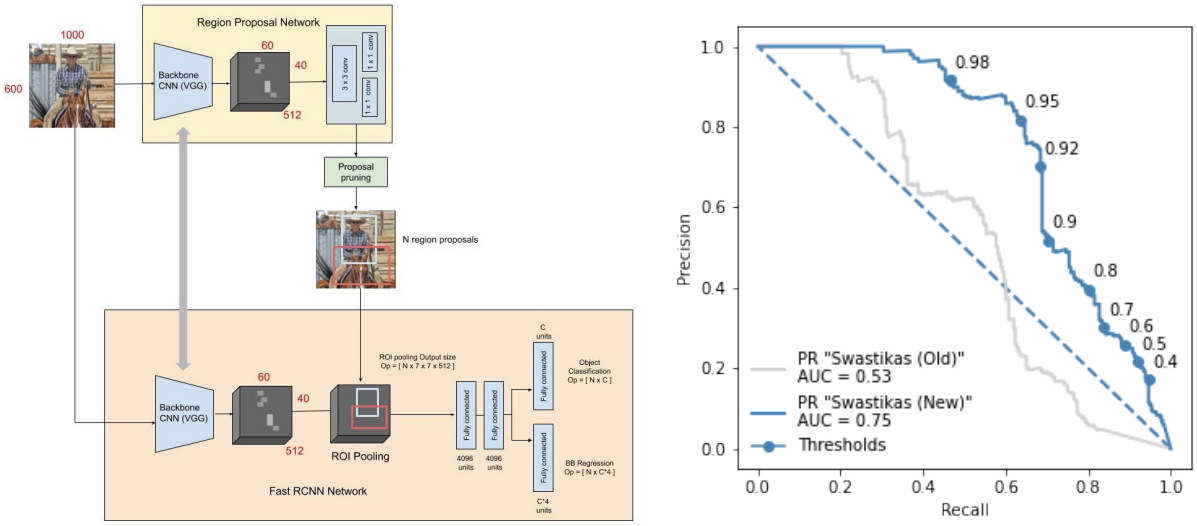
Tab. 1: Anzahl der annotierten Instanzen pro Annotationsklasse.

Mit diesem Dokument wurden die angestrebten Ergebnisse der Teilarbeitspakete AP 1.1, AP 2.1, AP 2.2 und AP 3.1 vollumfänglich erfüllt.

Klassifizierung von Hassrede und Bewertung der strafrechtlichen Relevanz

Der in dem oben genannten Anforderungsdokument beschriebene Bedarf an einer automatischen Klassifizierung von Hakenkreuzen als politisches Symbol des Nationalsozialismus in Bild- und Videodaten wurde wie folgt adressiert. Zunächst wurden geeignete Trainingsdaten erhoben. Hierzu wurden frei verfügbare Bilder (Creative Commons Lizenz) des Bilddatendienstes Flickr [2], die unter dem Suchbegriff “Swastika” zu finden waren heruntergeladen und vorsortiert. Ergänzt wurde dieser Datensatz durch vom BKA bereitgestellten Daten. Insgesamt wurden 1024 Bilder in den finalen Trainingsdatensatz aufgenommen. Zur Verbesserung der Vorhersagegüte wurde zusätzlich ein Datensatz mit 307 Screenshots von Postings in Sozialen Medien die keine Swastikas enthalten hinzugefügt, da diese Art von Daten einem der wichtigsten im Projekt betrachteten polizeilichen Anwendungsfällen entsprechen. Für die anschließende Annotation wurde zunächst deduktiv ein Codebook erstellt [3]. Hierbei wurde insbesondere getrennt nach kulturellem Kontext (rechte Ideologie “Rechts”, religiöser Kontext “Religiös”, antifaschistischer Kontext “Antifa”, uneindeutiger Kontext “Uneindeutiger Kontext”) der abgebildeten Symbole, sowie unvollständig abgebildete Symbole annotiert. Die einzelnen Annotationsklassen zusammen mit einigen typischen Beispielen sind in Abbildung 1 zu sehen. Insgesamt wurden 2873 Instanzen annotiert. Die Verteilung der Instanzen auf die verschiedenen Annotationsklassen ist in Tabelle 1 zu sehen.

Nach eingehender Literaturrecherche wurde als Basismodell für die Bildklassifizierung “Faster RCNN” [4] ausgewählt. Dieses wurde auf dem Coco Datensatz (Common Objects in Context) vortrainiert, welcher insgesamt 120.000 Bilder getrennt nach 80 Kategorien enthält [4]. Das Modell besteht aus zwei Modulen (siehe Abbildung 2). Ein Modul basiert auf einem Convolutional Neural



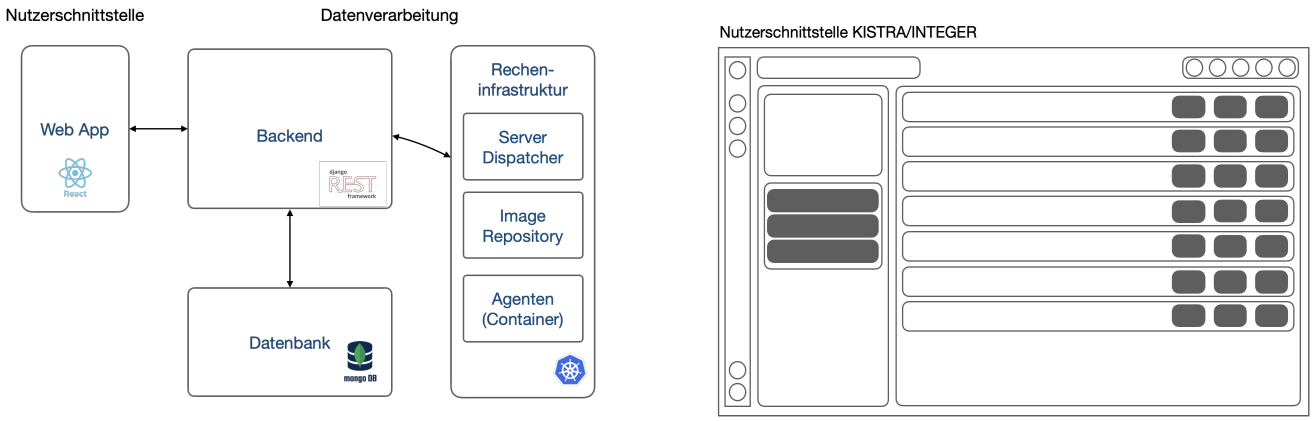
(a) (b)

Abb. 2: (a) Schematischer Aufbau des für die Bildklassifizierung verwendeten Modells (Faster RCNN). Zu erkennen sind die beiden getrennten Module zur Objektklassifizierung und zum Auffinden der Begrenzungsbox. (b) Receiver-Operating-Characteristic des finalen Modells (blaue Kurve) zusammen mit Schwellenwerten. Es wird eine Fläche unter der Kurve von 0.75 erreicht.

Network (CNN), welches für die eigentliche Bildklassifizierung zuständig ist. Ein zweites Modul identifiziert schließlich den Bereich im Bild, welcher mit hoher Wahrscheinlichkeit den größten Einfluss auf die Klassifizierung hat und markiert diesen mit einer Begrenzungsbox. Abschließend wurde die Güte der Vorhersagen evaluiert. In der finalen Ausprägung des Modells weist die entsprechende Receiver-Operating-Characteristic (ROC) eine Fläche unter der Kurve (Area Under Curve, AUC) von 0.75 auf und entspricht damit der Zielvorgabe des Anforderungsdokuments (siehe Abbildung 2). Der entwickelte Bildklassifizierer stellt somit den Beitrag von MIL zu den Teilarbeitspaketen AP 2.3 und 2.4 dar. Die Klassifizierung von weiteren verbotenen Symbolen wurde im Rahmen des Projektes nicht zuletzt wegen unzureichender Trainingsdaten nicht weiter verfolgt. In Zukunft soll dieser Bedarf aber durch einen Few-Shot-Learning-Ansatz adressiert werden, der mit entsprechend wenig Datenmaterial auskommt.

Technische Zusammenführung/Erstellung eines Frameworks und Evaluierung

Wie bereits eingangs beschrieben, basieren die technischen Lösungen, die im Rahmen von KISTRA-TSD erarbeitet werden, auf der KI-Recheninfrastruktur „MIL.ml Net“ und der OSINT-Analyseplattform „INspectre“. INspectre ist ein Rahmenwerk zur Erhebung und Analyse von Daten aus Sozialen Medien welches sich als Webanwendung einfach aus einem Browser heraus bedienen lässt. Eine Übersicht über die Hauptkomponenten von INspectre ist in Abbildung 3 (a) zu sehen. Ein wesentliches Merkmal ist hier die Trennung zwischen Nutzerschnittstelle



(a) (b)

Abb. 3: (a) Schematische Darstellung der Hauptkomponenten von INspectre sowie die Beziehungen zwischen diesen. (b) Schematische Darstellung der Erweiterung der Nutzerschnittstelle für die in KISTRA entwickelten Funktionen.

welche die Bedienung der Applikation ermöglicht, einem Backend zur Definition der Business-Logik und des Datenmodells und schließlich einer skalierbaren Recheninfrastruktur basierend auf Kubernetes, auf welcher die notwendigen Berechnungen für die einzelnen Datenanalysen ausgeführt werden. Diese Trennung hat unter anderem den Vorteil, dass die Recheninfrastruktur unabhängig von den anderen Komponenten in leistungsfähigen, ggf. mit spezieller Hardware wie Grafikprozessoren (GPUs) ausgestatteten Cloud-Umgebungen betrieben werden kann.

Zur Visualisierung und Interaktion mit den in KISTRA entwickelten Funktionen wurden der bestehenden Nutzeroberfläche weitere Elemente hinzugefügt (siehe auch Abbildung 3 (b)). Die in KISTRA entwickelten Funktionen sind also in einer Weise in INspectre integriert, dass sie sich nahtlos zusammen mit den bereits existierenden Funktionen nutzen lassen.

Eine schematische Darstellung der Recheninfrastruktur ist in Abbildung 4 zu sehen. Ein wesentliches Merkmal ist hier die Trennung der einzelnen Analysekomponenten in unabhängige, sogenannte Agenten. Auf diese Weise kann die Verarbeitungsgeschwindigkeit je nach Datenaufkommen und verfügbaren Ressourcen einfach durch Replikation der entsprechenden Agenten-Container im Kubernetes-Cluster gesteigert werden. Das erwartete Skalierungsverhalten wurde durch Benchmarktests des Gesamtdemonstrator auf einer Referenzinfrastruktur überprüft und bestätigt (siehe Abbildung 5 (a)). Der in der Anforderungsbeschreibung angestrebte Zieldurchsatz konnte erreicht werden. Als ein weiteres wesentliches Element des Gesamtdemonstrators wurden Mechanismen zur Überwachung des Betriebes der Applikation installiert. Ein Screenshot des Dashboards der hierzu verwendeten Metriken ist in Abbildung 5 (b) zu sehen.

Eine Dokumentation der Schnittstellen und des Software-Ökosystems von INspectre wurde mit den Forschungspartnern geteilt und die von diesen entwickelten Modelle bzw. Teildemonstratoren im Rahmen von Arbeitspaket 5 in den Gesamtdemonstrator integriert. Eine Übersicht der in den Gesamtdemonstrator integrierten Klassifizierer ist in Tabelle 2 zu sehen. Zur Darstellung und Interaktion mit den Modellvorhersagen wurden dem Framework neue Elemente der

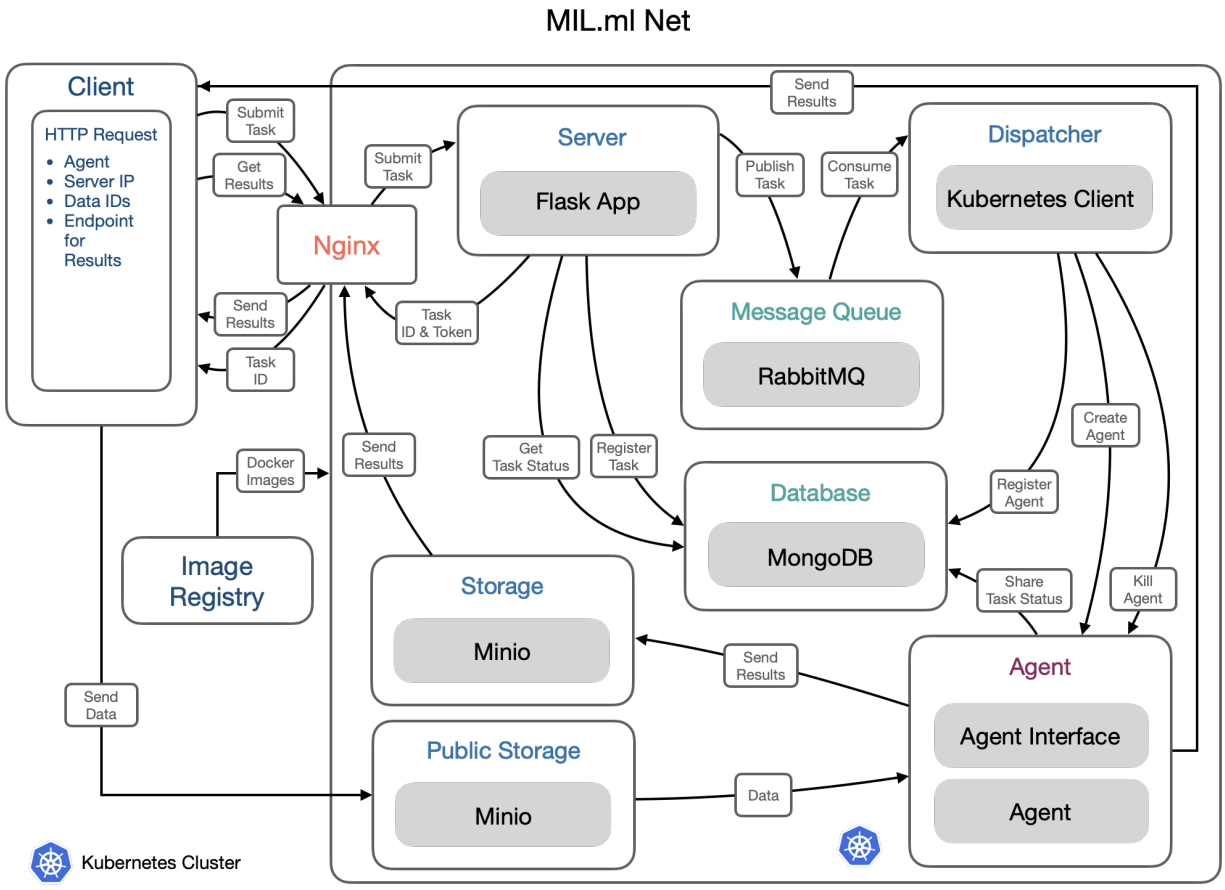
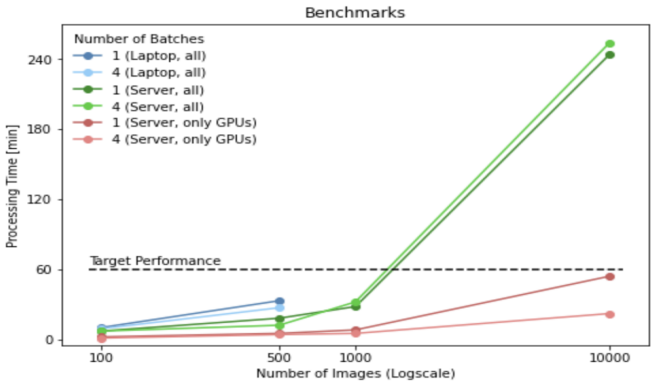
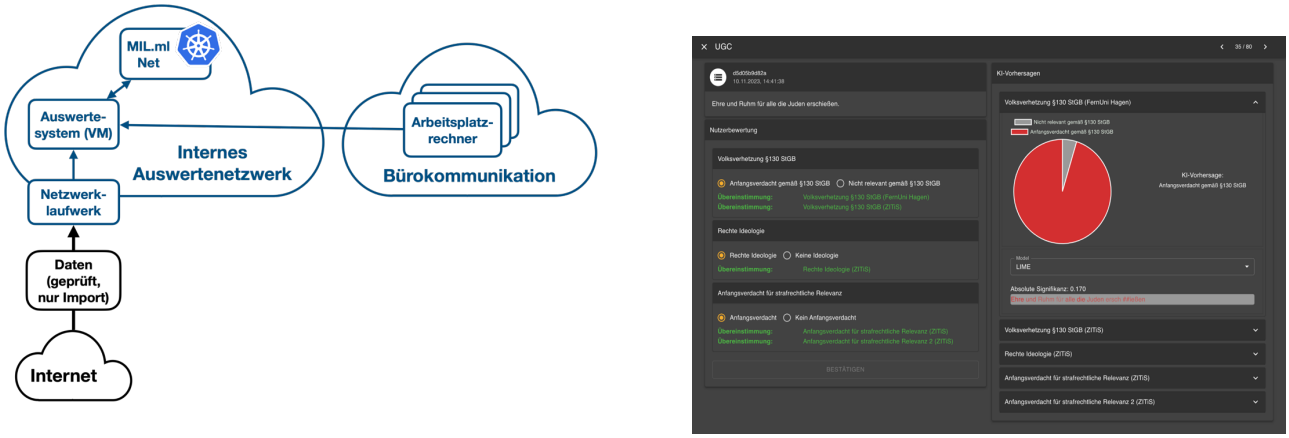


Abb. 4: Schematische Darstellung der Recheninfrastruktur MIL.ml Net des Gesamtdemonstrators. Die einzelnen Analysekomponenten sind in unabhängige, sogenannte Agenten organisiert.



(a) (b)

Abb 5: (a) Ergebnisse der durchgeführten Benchmark-Tests. Die in der Anforderungsdokumentation beschriebene Zielperformance (gestrichelte Linie) wurde erreicht (rote Kurven). (b) Screenshot des Dashboards der verschiedenen Metriken zur Überwachung des Betriebes der Applikation.



(a) (b)

Abb 6: (a) Schematische Darstellung der Integration des Gesamtdemonstrators in eine bestehende BOS-Infrastruktur. (b) Nutzerschnittstellenelemente zur Bewertung von Postings aus Sozialen Medien. Übereinstimmung zwischen der Bewertung durch die NutzerInnen und der KI-Modelle werden angezeigt sowie Erklärungen zu den Modellbewertungen.

Nutzerschnittstelle hinzugefügt. Diese wurden insbesondere gemäß eines innerhalb des Projektes entwickelten Arbeitsflusses zur Verminderung der Gefahr von Scheinprüfungen durch die NutzerInnen ausgewählt und angeordnet. So werden beispielsweise in der Konfiguration mit der geringstmöglichen Beeinflussung durch die KI die Vorhersagen der Modelle erst angezeigt, nachdem die NutzerInnen ihre Bewertung vorgenommen haben. Anschließend wird lediglich eine Übereinstimmung oder Abweichung von der Bewertung der KI-Modelle angezeigt. Es bleibt den NutzerInnen überlassen, ob sie ihre Entscheidung dann beibehalten oder revidieren möchten. Zusätzlich zu den Modellvorhersagen werden auch Erklärungen in Form von farblich hervorgehobenen Worten bzw. Wortbestandteilen, die den größten Einfluss auf die Modellentscheidung hatten, mitgeliefert. Ein Screenshot mit den entsprechenden Komponenten der Nutzerschnittstelle ist in Abbildung 6 (b) zu sehen. Es wurde ebenfalls aufgezeigt, wie jedem integrierten Modell weitere Informationen, beispielsweise bezüglich einer ggf. bestehenden Angreifbarkeit, wie sie in Arbeitspaket 4 untersucht wurde, beigefügt werden können (siehe auch Abbildung 7). Ein weiterer Fokus der Arbeit am Gesamtdemonstrator waren Maßnahmen zur Überwachung und Fehlererkennung des Anwendungsbetriebes sowie eine Entwicklung der einzelnen Softwarekomponenten welche einige Besonderheiten des Betriebes einer solchen Softwarelösung bei den BOS (On-Premise-Hosting, Schutzzonen, Datenhaltung, Mandantentrennung) berücksichtigt. Abbildung 6 (a) zeigt beispielhaft, wie eine solche Integration in eine bestehende BOS-Infrastruktur aussehen könnte. Der Gesamtdemonstrator läuft auf einer virtuellen Maschiene eines internen Auswertenetzwerkes. Dieses ist von der Bürokommunikationsumgebung sowie dem Internet durch verschiedene Sicherheitsmaßnahmen getrennt.

Die durch MIL erarbeiteten technischen Lösungen wurden während der Projektlaufzeit und abschließend einer Beurteilung durch Forschungspartner und Endanwender im Rahmen von mehreren Nutzertests unterzogen.

Bereich	Modell	Datentyp	Projekt	Verantwortlich	Status
Gesichter	Gesichter	Bilder/Videos	INTEGER	MIL	Integriert
Objekte	Waffen (Pistolen, Sturmgewehre)	Bilder/Videos	INTEGER		Integriert
Verbotene Symbole	§86a: Schwarzes Banner	Bilder/Videos	INTEGER		Integriert
	§86a: Swastikas	Bilder/Videos	KISTRA		Integriert
Straf-normen	§130 StGB (Volksverhetzung)	Text	KISTRA	FernUni Hagen	Integriert
	§241 StGB (Bedrohung)	Text	KISTRA		Wenige Beispiele
Phänomen-bereiche	PMK - Rechts	Text	KISTRA	ZITiS	Integriert
	PMK - Links	Text	KISTRA		Wenige Beispiele
	PMK - Religiöse Ideologie	Text	KISTRA		Wenige Beispiele
	PMK - Ausländische Ideologie	Text	KISTRA		Wenige Beispiele
	PMK - Nicht zuzuordnen	Text	KISTRA		Wenige Beispiele
	Keine politisch motivierte Ideologie	Text	KISTRA		Integriert
Themen-felder	Fremdenfeindliche Hasskriminalität	Text	KISTRA		Vorläufige Version
	Antisemitische Hasskriminalität	Text	KISTRA		Vorläufige Version
	Allgemeine Hasskriminalität	Text	KISTRA		Vorläufige Version
	Keine Hasskriminalität	Text	KISTRA		Vorläufige Version
Sonstige	Anfangsverdacht	Text	KISTRA		Integriert

Tab. 2: Übersicht der in den Gesamtdemonstrator integrierten Klassifizierer.

FaceScrub

Modell	Trainingsg…	Testgenaui…	Attacke	GAN	Attacken-G…	Feature-Di…	FID-…
ResNeSt-101a	100%	93,88%	Plug & Play Attacken	FFHQ	89,56%	0.732	47,9
ResNet-152	100%	93,74%	Plug & Play Attacken	FFHQ	92,73%	0,7163	46,69
DenseNet-169	100%	95,54%	Plug & Play Attacken	FFHQ	95,13%	0,6841	46,92
ResNeSt-101	100%	93,88%	Plug & Play Attacken	MetFaces	75,04%	0,9787	88,66
ResNet-152	100%	93,74%	Plug & Play Attacken	MetFaces	73,07%	0,9660	68,54
DenseNet-169	100%	95,54%	Plug & Play Attacken	MetFaces	79,83%	0,9376	77,52

Zeilen pro Seite: 10 1-6 von 6

Abb. 7: Informationen zur potenziellen Angreifbarkeit von verschiedenen ausgewählten Modellen.

3 Verwendung der Zuwendung

Das Projekt war mit erheblichen technischen und wirtschaftlichen Risiken verbunden. Die Erforschung der Erkennung von Hassnachrichten kombinierte moderne Ansätze von Künstlicher Intelligenz und Machine Learning mit phänomen- und länderspezifischen Anforderungen. Diese Kombination machte das Projekt besonders risikoreich.

3.1 Überblick über den zahlenmäßigen Nachweis

Für das Vorhaben und dessen Realisierung wurden insgesamt 584.199,77 EUR aufgewendet. Diese Aufwände gliedert sich in erster Linie in Personalkosten in Höhe von 579.221,63 EUR.

3.2 Notwendigkeit und Angemessenheit der Projektarbeiten

Die geleisteten Projektarbeiten waren notwendig und angemessen, da sie Personal mit hochqualifizierter IT-Expertise, fortgeschrittenem Wissen in Künstlicher Intelligenz und tiefem Verständnis für den Bereich Hassnachrichten erforderten. Diese Kombination von Fähigkeiten war entscheidend, um die technischen Herausforderungen zu meistern und präzise Ergebnisse zu erzielen. Ohne diese spezialisierten Kenntnisse wäre die erfolgreiche Durchführung des Projekts nicht möglich gewesen.

3.3 Voraussichtlicher Nutzen

Verbesserte Algorithmen: Durch die Erforschung der Erkennung von Hassnachrichten mittels KI können neue, präzisere Algorithmen entwickelt werden. Diese Algorithmen verbessern nicht nur die Erkennung von Hassnachrichten, sondern können auch in anderen Bereichen der Sprach- und Bildverarbeitung angewendet werden. Die im KISTRA Vorhaben gewonnenen Erkenntnisse in diesem Bereich wurden auch wissenschaftlich veröffentlicht.

Erweiterung des Wissens im Bereich Hassnachrichten und deren Bekämpfung: Die Studie trägt zur Erweiterung des wissenschaftlichen Verständnisses bei, wie Hassnachrichten entstehen und verbreitet werden. Dies ermöglicht tiefere Einblicke in die psychologischen und soziologischen Aspekte von Online-Kommunikation und trägt zur Entwicklung von präventiven Maßnahmen bei.

3.4 Relevante F&E-Ergebnisse Dritter

Während der Durchführung des Vorhabens wurden dem Zuwendungsempfänger relevante Forschungs- und Entwicklungsergebnisse von Dritten entsprechend beobachtet und im Rahmen der wissenschaftlichen Aufarbeitung berücksichtigt. Andere Stellen erzielten relevante Fortschritte in der Anwendung von Künstlicher Intelligenz zur Erkennung von Hassnachrichten, insbesondere in Bezug auf verbesserte

Algorithmen und Datenanalyse-Methoden. Diese Erkenntnisse flossen in die eigenen Arbeiten ein und trugen zur Optimierung der Forschung im Projekt bei.

3.5 Veröffentlichungen

Lenaršič, M. (2022, 08. Juni) AI model building for data analysis in LEAs: A practical example of using AI enabled OSINT solution for data analysis in law enforcement [Konferenzbeitrag].CEPOL Research & Science Conference 2022 MRU, Vilnius, Litauen.

Taing, S. (2023, 23./24. Mai) Die Zukunft von OSINT: Ein Ausblick auf das polizeiliche Internet- Monitoring im Jahr 2030 [Konferenzbeitrag]. 10. Internationales Symposium Neue Technologien, Stuttgart, Deutschland.

4 Literaturverzeichnis

[1] „Technische Anforderungen KI-Module/Gesamtdemonstrator“, Munich Innovation Labs, 2023

[2] https://www.flickr.com

[3] “Codebook Hakenkreuze als politisches Symbol des Nationalsozialismus”, Munich Innovation Labs, 2023

[4] https://towardsdatascience.com/faster-r-cnn-for-object-detection-a-technical-summary-474c5b857b46

5 Abkürzungsverzeichnis

AP	Arbeitspaket
BKA	Bundeskriminalamt
BOS	Behörden und Organisationen mit Sicherheitsaufgaben
CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
JGU	Johannes-Gutenberg Universität Mainz
KI	Künstliche Intelligenz
KISTRA	Projektname; ‘Einsatz von KI zur Früherkennung von Straftaten’
KISTRA-TSD	Teilvorhaben; ‘Technischer Software-Demonstrator’
LMU	Ludwig-Maximilians-Universität München
MIL	Munich Innovation Labs GmbH
RWTH	Rheinisch-Westfälische Technische Hochschule
TUB	Technische Universität Berlin
TUD	Technische Universität Darmstadt
UDE	Universität Duisburg-Essen
ZITiS	Zentrale Stelle für Informationstechnik im Sicherheitsbereich

6 Abbildungs- und Tabellenverzeichnis

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht

3. Titel

Einsatz von **KI** zur Früherkennung von **Stra**ftaten (**KISTRA**): Technischer Software-Demonstrator (KISTRA-TSD)

4. Autor(en) [Name(n), Vorname(n)]	5. Abschlussdatum des Vorhabens 31.12.2023
Taing, Stefan Uhlenbrock, Mathias	6. Veröffentlichungsdatum 28.06.2024
	7. Form der Publikation

8. Durchführende Institution(en) (Name, Adresse)	9. Ber. Nr. Durchführende Institution
Munich Innovation Labs GmbH Pettenkoferstr. 24 80336 München	10. Förderkennzeichen 13N15345
	11. Seitenzahl

12. Fördernde Institution (Name, Adresse)	13. Literaturangaben
Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	14. Tabellen
	15. Abbildungen

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)

18. Kurzfassung

Das Teilvorhaben KISTRA-TSD untersuchte den Einsatz von KI zur Früherkennung und Verfolgung von Straftaten, insbesondere Hasskriminalität, durch polizeiliche AnwenderInnen, mit dem Ziel der Entwicklung technischer Lösungen zur Klassifizierung von Hassrede sowie Bild- und Videodaten, um die polizeiliche Auswertung effizienter zu gestalten. Ein besonderer Schwerpunkt lag auf dem rechtlichen und ethischen Kontext der Datenverarbeitung durch Sicherheitsbehörden. Basierend auf dem KI-Framework „MIL.ml Net“ und der OSINT-Analyseplattform „INspectre“ wurden spezifische Anforderungen erfasst und dokumentiert, Trainingsdaten erhoben, Modelle trainiert und integriert. Der Gesamtdemonstrator beinhaltet angepasste Nutzerschnittstellen zur Minimierung von Fehlprüfungen und Berücksichtigung besonderer Betriebsanforderungen. Ein technisches Anforderungsdokument beschreibt die rechtlichen, ethischen und funktionalen Aspekte des Systems. Ein Bildklassifizierer zum Erkennen von Swastikas und weitere Modelle zur Erkennung von Hassrede wurden ebenfalls entwickelt und integriert. Die Nachvollziehbarkeit und Erklärbarkeit der Modellvorhersagen wurde schließlich durch Visualisierungen unterstützt. In NutzerInnentests wurden die KI-Modelle als hilfreich und die Umsetzung im Demonstrator als geeignet bewertet. Insgesamt zeigte das Projekt, dass die entwickelten Lösungen die polizeiliche Effizienz verbessern können, wobei rechtliche und ethische Rahmenbedingungen berücksichtigt werden.

19. Schlagwörter
Analyse Sozialer Netzwerke, Hasskriminalität, Künstliche Intelligenz (KI), Nutzeranforderungen, OSINT

20. Verlag	21. Preis

Document Control Sheet

1. ISBN or ISSN	2. type of document (e.g. report, publication) Final report

Use of AI in early detection of crime (**KISTRA**): Technical Software Demonstrator (**KISTRA-TSD**)

4. author(s) (family name, first name(s))	5. end of project 31.12.2023
Taing, Stefan Uhlenbrock, Mathias	6. publication date 28.06.2024
	7. form of publication
8. performing organization(s) (name, address) Munich Innovation Labs GmbH Pettenkoferstr. 24 80336 München	9. originator's report no.
	10. reference no. 13N15345
	11. no. of pages
12. sponsoring agency (name, address) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. no. of references
	14. no. of tables
	15. no. of figures

16. supplementary notes

17. presented at (title, place, date)

18. abstract

The subproject KISTRA-TSD investigated the use of AI for the early detection and tracking of crimes, particularly hate crimes, by police users. The aim was developing technical solutions for classifying hate speech as well as image and video data to make police analysis more efficient. A particular focus was placed on the legal and ethical context of data processing by security authorities. Based on the AI framework "MIL.ml Net" and the OSINT analysis platform "INspectre," specific requirements were collected and documented, training data was gathered, models were trained and integrated. The overall demonstrator includes customized user interfaces to minimize false positives and take into account special operational requirements. A technical requirements document describes the legal, ethical, and functional aspects of the system. An image classifier for recognizing swastikas and other models for detecting hate speech were also developed and integrated. The traceability and explainability of the model predictions were ultimately supported by visualizations. In user tests, the AI models were found to be helpful and the implementation in the demonstrator was deemed suitable. Overall, the project demonstrated that the developed solutions can improve police efficiency while taking into account legal and ethical frameworks.

19. keywords
Artificial Intelligence (AI), Hate Crime, OSINT, Social Network Analysis, User Requirements

20. publisher	21. price