CENTER FOR DIGITAL
SAFETY & SECURITY

AUSTRIAN INSTITUTE
OF TECHNOLOGY

FORSCHUNGSSCHWERPUNKT DESINFORMATIONSDETEKTION

Projektübergreifende, Intersektoriale Zusammenhänge
und Koordination

DESINFORMATIONSERKENNUNG

PROJEKTLINIE

STUDIE ZUR ERKENNUNG VON DESINFORMATION

- **Überblick über technologische Optionen zur Bekämpfung von Desinformation**
- Erster Tech-Pilot

defalsif-ai

- Entwicklung einer umfangreichen **Media-Forensik-Toolbox**
- Audiovisuelle Forensik zur Erleichterung von Faktenchecks
- **Erkennung von Audiomanipulationen**
- **Erkennung von Bild-/Videomanipulationen**
- **Deep Fake** Erkennung
- **Textinhaltsanalyse** (z. B. Schreib-/ Berichterstattungsstil, Tatsachen-behauptungen, Propaganda)

STARLIGHT

- **Einfach einsetzbare Tools für LEAs**
- Erkennung von Bildmanipulationen
- Analyse von Textinhalten

DesinFact

- Netzwerk/Graphenanalyse von Desinformationsnetzwerken
- Fokus auf Vertrauenswürdigkeit
- Fokus auf Präsentation und Interaktion
- Verbesserung der Qualität von AI-Modellen

EUCINF

- EDF Projekt
- Lösung für hybriden Bedrohungen in verschiedenen Szenarien
- Entwicklung einer Toolbox an KI-Tools zur Bekämpfung von Desinformation und hybriden Bedrohungen

2019 2020 2021 2022 2023 2024

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

- Analyse von Social-Media-Kanälen im Hinblick auf **Hate Speech** und **extremistische Inhalte**
- Ansätze zur Bekämpfung der **Infodemie** (Unterstützung bei der Bewältigung der Informationsflut)
- Analyse von **Hate Speech** und **toxischen Inhalten** (z. B. Sexismus, Toxizität, Diskriminierung)
- Analyse **extremistischer Inhalte** (z. B. politische, religiöse, kriminelle Relevanz)

German-Austrian Digital Media Observatory GADMO

- Aufspüren und Analysieren von Desinformationskampagnen
- Unterstützung von Mainstream-Medien, lokalen Medien und Behörden bei der Aufdeckung von Desinformationskampagnen
- Organisation von Aktivitäten zur Medienkompetenz auf nationaler oder multinationaler Ebene
- Unterstützung der nationalen Behörden bei der Kontrolle von Richtlinien für Online-Plattformen und des digitalen Medien-Ökosystems

HYBRIS

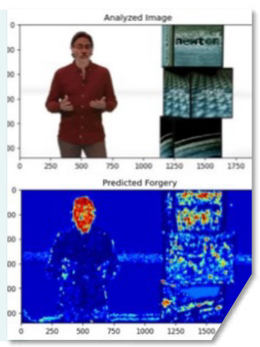
- Identifizierung und Analyse von **hybriden Bedrohungen**
- **Large Scale** Desinformations-Trendanalyse
- **Hochleistungsfähige Machine Learning Module**
- Erkennung von **Narrativen**
- Verbesserte **infodemische Unterstützung**

Defame Fakes

- Erkennung und Analyse von Deepfakes
- Konzept zur Erkennung von Deepfakes in Echtzeit
- Digitaler Bild- und Videodatensatz
- Modellübergreifende Inhaltsanalyse
- Kontextanalyse unter Verwendung von Open-Source-Daten
- Teil-automatisierte Softwaretools
- GSK und rechtliche Analyse der Regulierung
- Nationale Umsetzung des Deepfakes-Actionplans

ZIELSETZUNG

- Erkennung von Manipulationen in Medien
- Erkennung von künstlich erzeugten Medien und Deepfakes
- Methoden zur Nachvollziehbarkeit und v Beweisbarkeit beim Einsatz von KI-Methoden zur Erkennung Fake-News
- Analyse der rechtlichen Situation und der Möglichkeiten um z.B. gegen Deepfakes vorzugehen.



Erkennen von Deep Fake Manipulatioinen in Videos

PROJECT LINE DISINFORMATION DETECTION

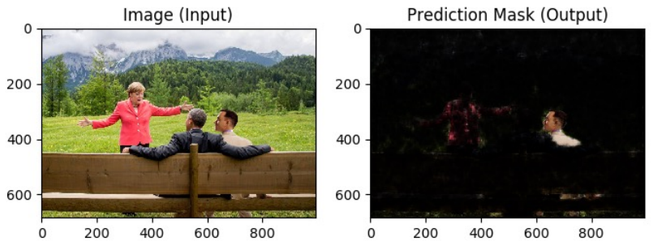
STUDIE → DEFALSIF-AI → RADAR → HYBRIS → .... → FAKE NEWS FRÜHWARNSYSTEM

AUFGABEN UND BEDROHUNGSBEREICH				
Studie zu Bedrohungslagen Technologien, Gegenmaßnahmen, Investionsstrategie, Empfehlungskatalog	Detektion von Desinformation, audiovisueller Medien-Manipulation, Analyse von Textinhalten	Erkennung von Hass im Netz, Radikalisierung, demokratiegefährdender Inhalte. Gefahrenpotential Analyse	Erkennung von Desinformations-Kampagnen in Big Data Strömen. Resilienz gegenüber Hybrider Bedrohungen	Multi-Stake-holder Plattform: "Wetterdienst" für Fake News Trends. Wissensbasis zu Desinformation.
ANWENDUNGSBEREICHE				
Einzelne Dateien	Einzelne Dateien Web-URLs	Einzelne Social Media Kanäle Konfiszierte Festplatte, Mobiltelefone	Vielzahl unterschiedlicher Social Media Kanäle Unterschiedliche heterogene Quellen	Unbegrenzte Zahl an heterogenen Kanälen, Quellen und Inhalten
ANALYSE UND DETEKTION				
erster Deep Fake Erkennungs-Prototyp	Manipulationen in Bild und Ton Deep Fakes Umfangreiche Text Analysen	Hate Speech Text-Analyse: Sexismus, Antisemitismus, Radikalismus Radikale Symbolik	Fake News Narrative Themen-Erkennung / Trend-Analyse Automatische Zusammenfassung	Trans-nationale / Quellen-übergreifende Trend-Analyse Cluster-Analyse
VERSTEHEN / WISSENSGEWINNUNG / TRENDS ERKENNEN				
Überblick über Bedrohungslagen und technische Möglichkeiten	Bild- und Audio-Manipulationen erkennen und erklären	Überblick über Themen und Inhalte in größeren Kanälen gewinnen	Fake News Narrative (Einsprachig) Lokale Fake News Trends	Multi-Linguale Narrativ-Fusion Globale Fake News Trends
ERGEBNISSE				
Berichte Empfehlungskatalog	Analyse-Plattform Medienforensik	Analyse-Plattform für Datenströme	Big Data / HPC Analyseplattform	Online-Plattform für Fake News Trends

KI-BASIERTE FACT-CHECKING TOOLS

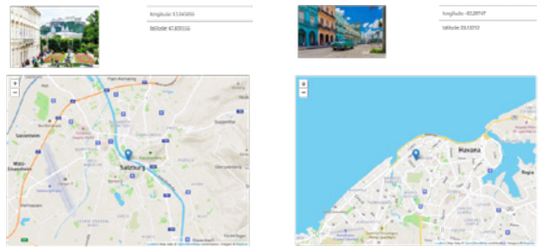
ANSATZ

- Bereitstellen von Werkzeugen zur Unterstützung von Fact-Checkern
- Medienforensiche Erkennung von Manipulationen
- Erkennen von synthetischen Inhalten

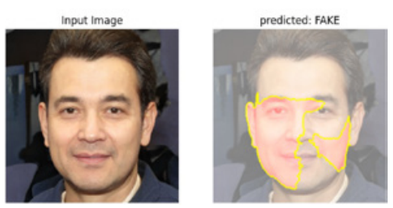


ERKENNUNG VON BILDMANIPULATION
KI basierte Erkennung ob an einem Bild etwas manipuliert – eingefügt / gelöscht - wurde. Verständliche Darstellung der Analyse Ergebnisse. Im Bild rechts wird angezeigt, was im Bild links ergänzt wurde.

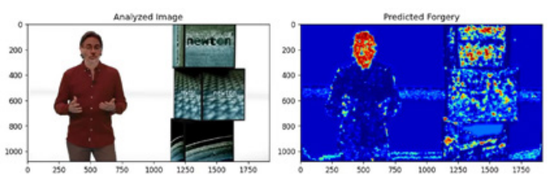
ERKENNEN DES AUFNAHMEORTS
Oft ist es wichtig zu überprüfen, ob ein Bild tatsächlich an dem angegebenen Ort aufgenommen wurde. Hierfür wurden Modelle entwickelt, welche den Ort der Aufnahme bestimmen können. Diese Methode funktioniert an bekannten Orten sehr gut, aber auch im freien Gelände mit einer Genauigkeit von bis zu 100Km Abweichung.



ERKENNEN VON FAKE PROFIL-FOTOS
Fake Profile in Sozialen Medien werden ein immer größeres Problem. Mittels generativen Modellen können immer bessere Fake Profil-Bilder erstellt werden. Unser neuronales Netz wurde mit 125.000 m Bildern aus verschiedenen Quellen trainiert und erzielt auf Benchmark-Datensätzen eine Korrektheit von 95-99,8 %.



ERKENNEN VON DEEP FAKES



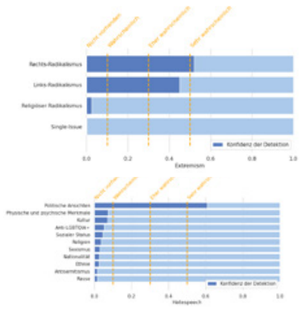
TEXT INHALTS-ANALYSE

Herausforderung

- Direkte Erkennung von Desinformation oft kaum möglich
- Benötigt breites Allgemeinwissen (in KI nicht vorhanden)

Ansatz

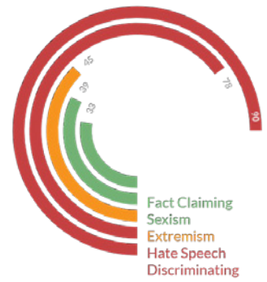
- Bestimmung mehrerer relevanter Inhaltsbeschreibungen und Merkmale
- Darstellung mittels Information Nutrition Labels
- Multi-Modale Fusion der Merkmale zu einer Gesamtbeurteilung
- hinsichtlich des (des-) Informationsgehaltes.



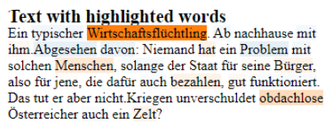
KI MODELLE zur Inhaltsbeschreibung

- Jedes Inhalts-Merkmal wird durch ein eigenes KI Modul aus dem online Daten abgeleitet.
- Beschreibung des (des-) Informationsgehaltes.
- Portfolio an KI Modulen über mehrere Projekte hinweg entwickelt (siehe Tabelle)

Verständliche Darstellung
Eine klare und übersichtliche Darstellung von Ergebnissen und Informationen steht ebenso im Fokus der Forschungstätigkeiten. Hierfür werden neue Ansätze zur Visualisierung erforscht.



Information Nutrition Labels
stellen beschreiben den Inhalt von Dokumenten oder Online Artikeln in einer übersichtlichen Form dar. Benutzer erhalten eine schnelle Einschätzung des Informationsgehalts.

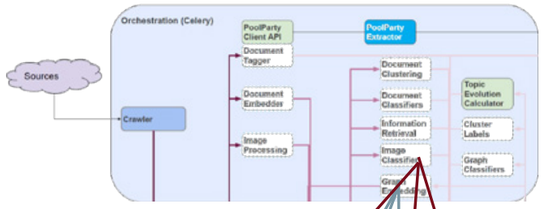


Erklärbarkeit von KI
Erklärbarkeit und einfache Verständlichkeit sind zentrale Anforderungen an KI Module. Der Benutzer muss stets Entscheidungen und Einschätzungen der KI interpretieren können

NAME	ERKANNTE INHALTE	SPRACHE	DOMÄNE	KATEGORIE BEISPIELE
Fake News	Direkt Erkennung von Fake News	Englisch	Soziale Netzwerke	Ja / Nein
Hassrede	Hass gegen Gruppen oder Individuen	Multi-ling	Soziale Netzwerke Diskussionsforen	Ja / Nein
Extremismus	Extremisitische Inhalte	Deutsch	Soziale Netzwerke Artikel	Rechts-, Links-, Religiös oder Single-Issue Extremismus
Toxizität	Giftige, beleidigende Inhalte, Kommentare, hasserfüllte Sprache	Deutsch	Soziale Netzwerke	Ja / Nein
Faktuelle Behauptungen	Wurde faktuell Behauptet?	Multi-ling	Soziale Netzwerke	Ja / Nein
Ansprechende Inhalte	Ansprechende, positive, diskussionsfördernde, Sprache	Deutsch	Soziale Netzwerke Artikel	Ja / Nein
Sentimentalität	Sentiment, Empfindung, Gefühl	Deutsch	Artikel	Positiv, Negativ
Berichtsstil	Berichtstil eines Artikels	Deutsch	Artikel	Verschwörungstheorie, Clickbait
Schreibstil	Schreibistil eines Artikels	Deutsch	Artikel	Polarisieren, Übertreibung
Diskriminierung	Ist eine Aussage Diskriminierend?	Deutsch	Soziale Netzwerke	Ethnie, Sozialer Status
Strafrechtliche Relevanz	Ist eine Aussage Kriminell?	Deutsch	Soziale Netzwerke	Verhetzung, Beleidigung
Sexismus	Diverse Kategorien von Sexismus	Englisch	Soziale Netzwerke	Misogynie, Sexuelle Gewalt

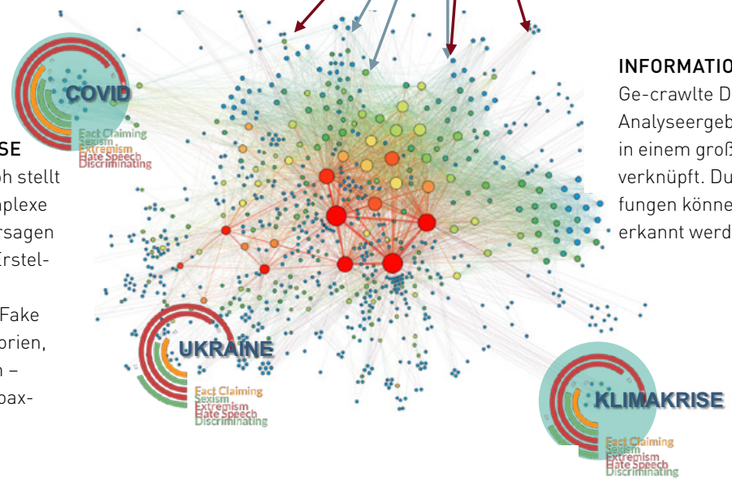
FAKE NEWS TREND ANALYSE

PRIVACY AWARE DATA AKQUISITION
Intelligente Crawler für unterschiedliche soziale Netzwerke und Plattformen, welche unter Berücksichtigung des Datenschutzes, relevante Daten automatisiert beziehen.



KOMPLEXE KI-PIPELINES
Desinformation ist komplex und benötigt viele spezifische KI-Module zur Erkennung. Jeder Artikel wird von einer Vielzahl von Modulen analysiert. Das effiziente Management solcher komplexen Pipelines benötigt optimale Planung und Ingeneursleistung.

WISSENSGRAPH ANALYSE
Der erstellte Wissensgraph stellt die Ausgangslage für Komplexe Analysen und Trendvorhersagen dar. Er kann auch für die Erstellung einer umfangreichen Wissens-Datenbank über Fake News, Verschwörungstheorien, etc. herangezogen werden – ähnlich wie bestehende Hoax-Email Datenbanken.

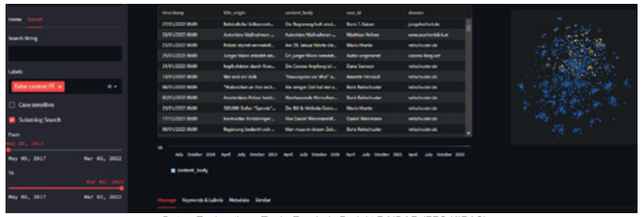


INFORMATIONS-VERNETZUNG
Ge-crawlte Daten werden mit Analyseergebnissen der KIModule in einem großen Wissensgraphen verknüpft. Durch diese Verknüpfungen können Zusammenhänge erkannt werden.

NETZWERK ANALYSE
Erkennen von Distributionswegen und Schlüssel-Aktoren in Desinformationsnetzwerken. Erkennung und Analyse von Echo-Chambers und Bot-Netzwerken.

GRAPH-KI ANALYSE
Graph Neural Networks sind der neueste Trend im Bereich Künstliche Intelligenz. Diese Vielversprechende Technologie ermöglicht es hochkomplexe Zusammenhänge zu modellieren und auszuwerten. Speziell für so komplexe und subjektive Aufgabenstellungen wie die Interpretation des (des-) Informationsgehaltes, stellen sie eine optimale Lösung dar, um die unterschiedlichen Datenformate (Text, Bild/Video, Ton, Beziehungen in Sozialen Netzwerken, etc.) miteinander zu Verknüpfen, bzw. um Verknüpfungen automatisiert zu erkennen.

ÜBERSICHTLICHE THEMEN-DARSTELLUNG
Themencluster mittels Information Nutrition Labels visualisiert. Rascher Überblick durch automatisch extrahierte Keywords und Kurz-Zusammenfassungen.


Daten-Explorations-Tool - Ergebnis Projekt RAIDAR (FFG KIRAS)

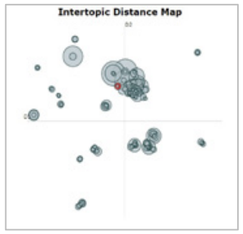

Trend Analyse in globalen Nachrichten –

INFODEMIC BEKÄMPFEN

ZU VIELE INFORMATIONEN ÜBER ZU VIELE KANÄLE
Infodemic beschreibt die Ohnmacht gegenüber der permanenten Flut an Nachrichten, in welcher es nicht mehr möglich ist, zu unterscheiden, ob etwas wahr oder falsch ist.

ANSATZ
- Inhalte automatisiert strukturieren
- Aus großen Mengen an Nachrichten, relevante Inhalte
- zusammenfassen
- Übersichtliche Informations-Visualisierung
- Beziehungen und Ähnlichkeiten darstellen

THEMEN ERKENNUNG
Automatisches Erkennen von Zusammenhängen, basierend auf Text-Ähnlichkeit und semantischer Analyse. Übersichtliche Darstellung von Themen-Clustern und deren Ähnlichkeiten. Hirarchische Gliederung in Sub-Themen.



INFODEMIC IS
"... an overabundance of information – some accurate and some not – that makes it hard for people to find trustworthy sources and reliable guidance when they need it"

DARSTELLUNG VON SEMANTISCHER ÄHNLICHKEIT
Berechnen und Darstellung von Ähnlichkeiten in Medienkollektionen – z.B. Bilder, Texte, Videos - damit Benutzer Zusammenhänge besser erkennen können.

SCHLÜSSELWORT ERKENNUNG
Automatisches Erkennen relevanter Schlüsselwörter. Ermöglichen eines schnellen Überblicks über Inhalte eines Artikels oder eines oder mehrer Social Media Kanäle.

AUTOMATISCHE KURZ-ZUSAMMENFASSUNG
Kurze Zusammenfassung eines oder mehrerer Artikel, um einen schnellen Überblick über geteilte Inhalte oder geführete Diskussionen zu erhalten.

KOOPERATIONSPARTNER

MINISTERIELLE KOOPERATIONEN

- Bundeskanzleramt
- Bundesministerium Landesverteidigung
- Bundesministerium Justiz
- Bundesministerium Inneres
- Bundesministerium Europäische und internationale Angelegenheiten

INSTITUTIONELLE KOOPERATIONEN

ORF · APA Austriapresseagentur · dpa Deutsche Presse-Agentur GmbH · AFP · CORRECTIV Recherchen für die Gesellschaft

FORSCHUNGS- UND INDUSTRIEPARTNERSCHAFTEN

LIquA · research institute · SCENOR The Science Crew · Donau-Universität Krems Universität für Weiterbildung · thinkers.ai Best in results · TU Wien Technische Universität Wien Vienna University of Technology · BOKU Universität für Bodenkultur Wien · enliteAI · Semantic Web Company · Artificial Researcher · Vienna Scientific Cluster · ATC · technische universität dortmund

FÖRDERPROGRAMME

FFG Forschung wirkt. · KIRAS Sicherheitsforschung · European Commission · Horizon 2020 European Union funding for Research & Innovation

DR. ALEXANDER SCHINDLER
Thematic Coordinator Datascience
Data Science & Artificial Intelligence
Center for Digital Safety & Security

AIT Austrian Institute of Technology
Giefinggasse 4 | 1210 Vienna | Austria
+43 664 8251454
alexander.schindler@ait.ac.at

DI. (FH) MARTIN BOYER
Senior Research Engineer
Sensing & Vision Solutions
Center for Digital Safety & Security

AIT Austrian Institute of Technology
Giefinggasse 4 | 1210 Vienna | Austria
+43 664 8251440
martin.boyer@ait.ac.at

MAG. MICHAEL MÜRLING
Marketing and Communications
Center for Digital Safety & Security

AIT Austrian Institute of Technology
Giefinggasse 4 | 1210 Vienna | Austria
T +43 50550-4126
michael.muerling@ait.ac.at