

What is Artificial Intelligence?

Academic Discipline / Computer Science



Dartmouth Conference (1956)

- Naming: „Artificial Intelligence“
- Multi-disciplinary
 - Philosophy (z.B. Descartes, Leibnitz)
 - Logic / Mathematics (z.B. Gödel)
 - Computer Science (z.B. Turing, von Neumann)
 - Psychology / Cognitive Science (Knowledge representation)
 - Biologie / Neuro-Wissenschaften (Konnektivismus, Neural Networks)
 - Evolution (Genetic Programming)

- “*Artificial Intelligence (AI) is the part of computer science concerned with **designing intelligent computer systems**, that is, systems that exhibit **characteristics we associate with intelligence in human behavior**” (Barr & Feigenbaum, 1981)*

- Understanding language
- Learning
- Reasoning
- solving problems

- **Scientific Goal:** To determine which ideas about knowledge representation, learning, rule systems, search, and so on, explain various sorts of real intelligence.

- **Engineering Goal:** To solve real world problems using AI techniques such as knowledge representation, learning, rule systems, search, etc.

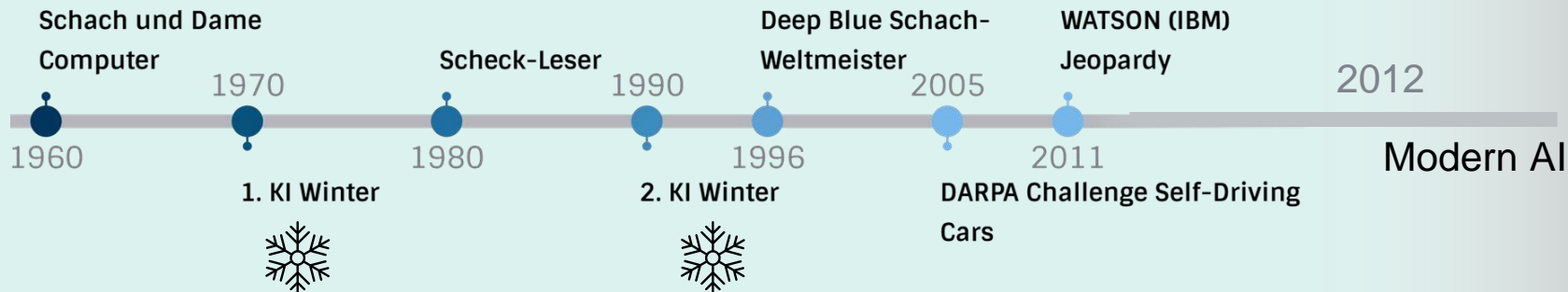
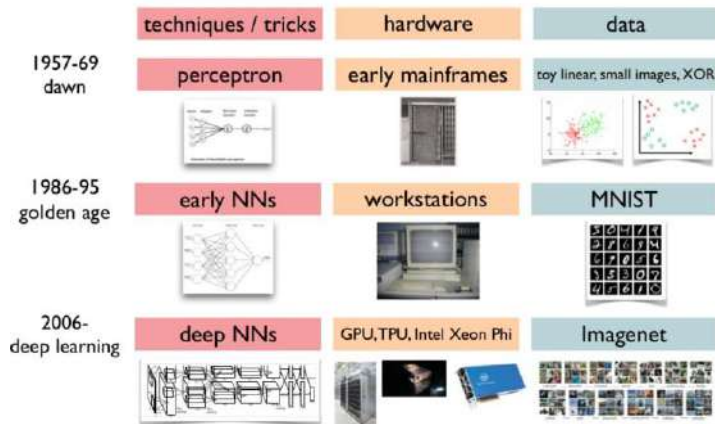
HISTORICAL OVERVIEW

High expectations

- Major investment from the military
- Utopian ideas

Poor performance

- Slow computers
- Small data sets / Expensive data storage
- Many problems not yet solved
- Too few "experts"



DEFINITION

ARTIFICIAL INTELLIGENCE

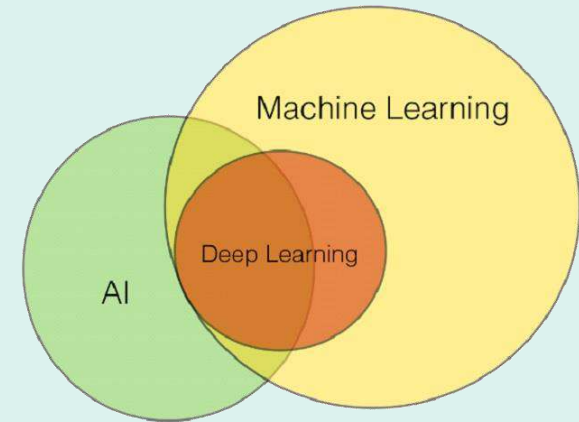
Table 1. AI domains and subdomains constituting one part of the operational definition of AI

AI taxonomy		
	AI domain	AI subdomain
Core	Reasoning	Knowledge representation
		Automated reasoning
		Common sense reasoning
	Planning	Planning and Scheduling
		Searching
		Optimisation
	Learning	Machine learning
Communication	Natural language processing	
Perception	Computer vision	
	Audio processing	
Transversal	Integration and Interaction	Multi-agent systems
		Robotics and Automation
		Connected and Automated vehicles
	Services	AI Services
	Ethics and Philosophy	AI Ethics
Philosophy of AI		

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). AI Watch Defining Artificial Intelligence. Publications Office of the European Union.

Machine Learning

- Methods which **leverage data** to **improve performance** on some set of **tasks**
- Statistical Machine Learning is dominant



DEEP LEARNING REQUIRES DATA

Learning by examples

- Deep Learning requires a lot of diverse data to recognize the relevant features of an object.
 - Different dog breeds, car brands, bird species, clothing, etc.

Overfitting

- If a model "sees" only red cars, it assumes that cars are red.
- If the training data is not diverse enough, it cannot generalize to "unseen" data ("it does not recognize blue cars")

Problems

- Bias: Annotator's bias, prejudices, stereotypes, etc. are reflected in the annotation / data sampling process
- Personal Data: Identifying information in text, image, sound.
 - → Anonymizing data, synthetic data
- Ethical implications: What data may be used (e.g., skin color, religion, sexual orientation, consumer behavior, etc.)?
 - → Ethics guidelines, legal guidelines



DATA SCIENCE

Science in dealing with data / handling data

- „Data Science“ originates in the 1960s
 - importance of statistical data analysis for an understanding of data was foreseen in an article¹ in 1962
- Gained increased importance with „Big Data“
- **Focus:**
 - Not on the data itself
 - rather on the way in which the data is
 - processed, prepared, analysed
 - translated into decisions
- **Data science is concerned with**
 - purpose-oriented data analysis
 - systematic generation of decision-making aids, tools and bases
 - to achieve competitive advantages

Conceptual Framework	Introduction to Data
Data Collection	Data Discovery and Collection
	Evaluating and Ensuring Quality of Data and Sources
Data Management	Data Organization
	Data Manipulation
	Data Conversion
	Metadata Creation and Use
	Data Curation, Security and Re-Use
	Data Preservation
Data Evaluation	Data Tools
	Basic Data Analytics
	Data Interpretation (Understanding Data)
	Identifying Problems Using
	Data Visualization
	Presenting Data (Verbally)
	Data Driven Decisions Making (DDDM)
Data Application	Critical Thinking
	Data Culture
	Data Ethics
	Data Citation
	Data Sharing
	Evaluating Decisions based on Data

Abbildung 1: Data-Literacy-Kompetenzen nach Ridsdale et al.

Data Literacy und Data Science Education: Digitale Kompetenzen in der Hochschulausbildung. Policy Paper der Präsidiums-Task-Force „Data Science“ der Gesellschaft für Informatik e.V. in Zusammenarbeit mit Vertretern der Deutschen Mathematiker-Vereinigung e.V., der Deutschen Physikalischen Gesellschaft e.V. und der Gesellschaft Deutscher Chemiker e.V.

ÖVP bremst Mietpreisbremse aus

POLITIK / 22.09.2023 • 17:47 Uhr / 7 Minuten Lesedort



Wolten wird abermals teurer. ÖVP und ÖRdne konnten sich auf keine Mietpreisbremse einigen. [https://www.vol.at](#)

Stattdessen kommt als Kompromiss eine einmalige Wohnbeihilfe von etwa 200 Euro. Wer darauf Anspruch hat, setzen die Bundesländer fest.



Julia Schilly
julia.schilly@vol.at

WIEN Es kommt keine Mietpreisbremse. Das ist seit Mittwoch klar.

Software Engineering for AI projects ...

... Requirements Engineering

- Complex
- runaway expectations

... SW Architecture

- Wide, complex and relative new field
- Best practices not yet widely incorporated into teaching

... SW Management Process

- Best practices not yet widely incorporated into teaching

... SW Maintenance

- Slow and difficult debugging



ÖVP bremst Mietpreisbremse aus

POLITIK | 22.09.2023 | 17:47 Uhr | 7 Minuten Lesedauer



Wolten wird abermals teurer. ÖVP und Örtliche könnten sich auf keine Mietpreisbremse einigen. | [Mehr zum Thema](#)

Stattdessen kommt als Kompromiss eine einmalige Wohnbeihilfe von etwa 200 Euro. Wer darauf Anspruch hat, setzen die Bundesländer fest.



WIEN Es kommt keine Mietpreisbremse. Das ist seit Mittwoch klar.

... AI Systems Risk Management

• Extremely difficult to calculate

- multiple inter-disciplinary impacts
- Competences often not covered by executing institutions (especially KMUs)
- Knowledge about AI System management requirements not consolidated (AT, EU)
 - → difficult to estimate vendor/outcome reliability

• Impact on society

- High to very high, e.g.
 - violation of personal rights
 - IT personel defining classification systems → becomes new reality/truth
 - Loss of trust in promising technology

• Impact on personal lifes

- High to very high
- E.g., damaged reputation, credits not granted

• Impact on hosting company

- High
- High visibility
- Damaged reputation
- Legal consequences

• Impact on developing company

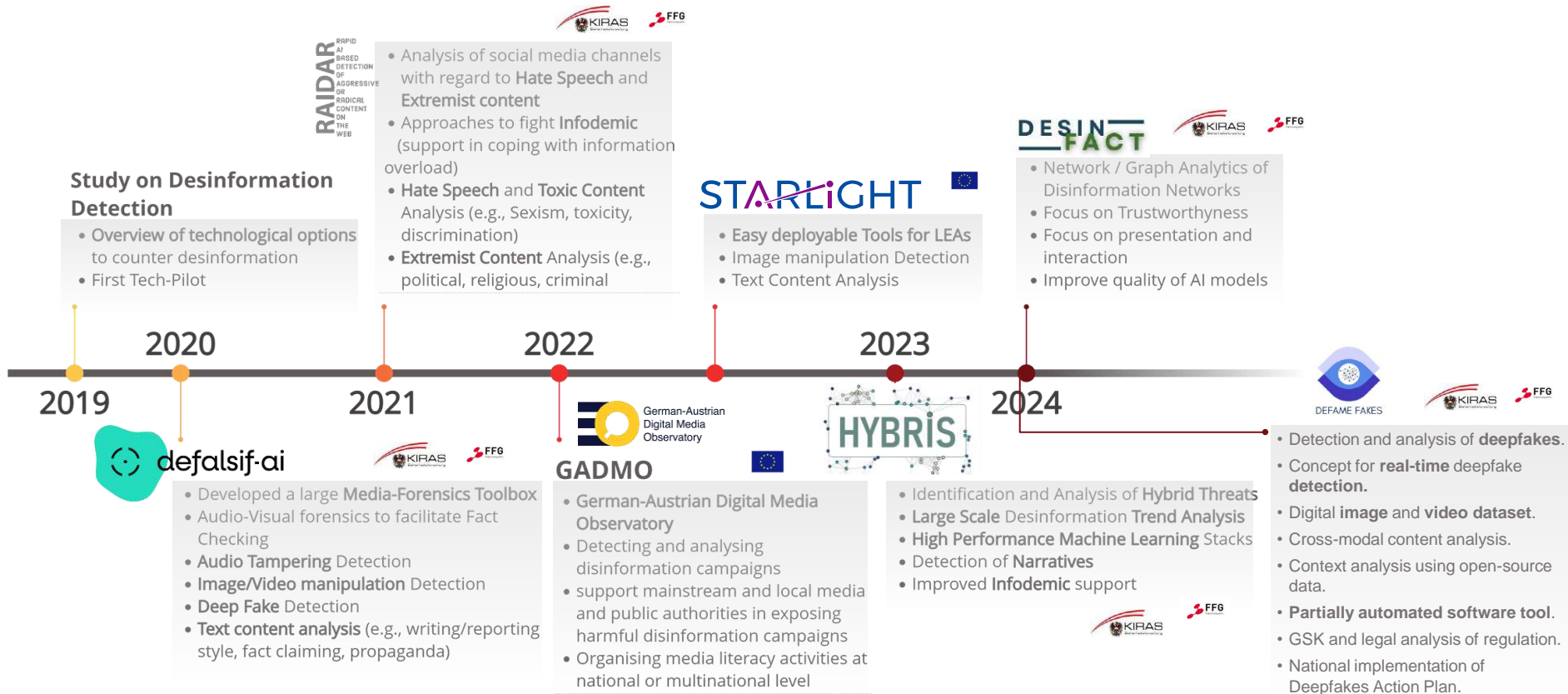
- High
- Damaged reputation
- Loss of funding, future comissioning
- Legal consequences

Projects in Civil Security



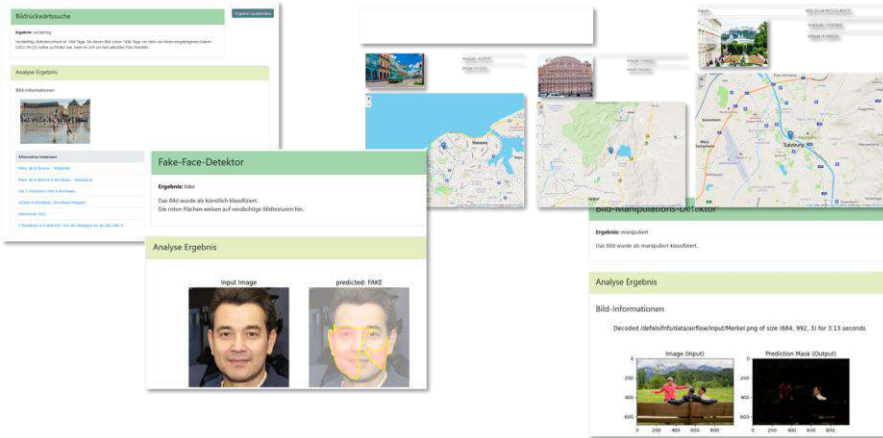
Disinformation Detection Research @ AIT

Projects



Visual Signals for the detection of Disinformation

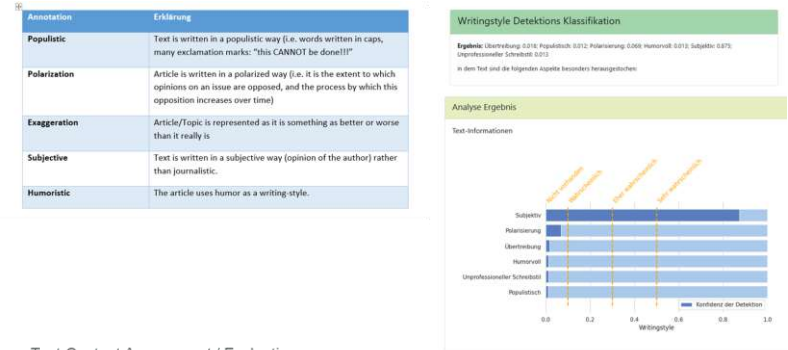
Text Content Assessment / Evaluation



The collage shows several user interfaces for disinformation detection tools:

- Bild-Geo-Location:** A map interface showing a location in Salzburg, Austria, with a street view image and a list of nearby points of interest.
- Fake-Face-Detector:** An interface showing an input image of a man's face and a predicted face mask with yellow highlights on the eyes and mouth area. The text below reads: "Eigenschaften: Das Bild wurde als künstlich klassifiziert. Die roten Flächen weisen auf verdächtige Bildbereiche hin."
- Image Manipulation Detector:** An interface showing an input image of a group of people sitting on a bench and a predicted mask (Output) with a bounding box around the image. The text below reads: "Analyse Ergebnis: Decoded: defalsif/bilderanalyse/inputs/Market.png of size 1084_962_31 for 3.13 seconds."

- Reverse-Image Search
- Image Geo-Location Estimation
- Fake-Face-Detector
- Image Manipulation Detector



The interface displays the following information:

Annotation	Erklärung
Populistic	Text is written in a populist way (i.e. words written in caps, many exclamation marks: "this CANNOT be done!!!")
Polarization	Article is written in a polarized way (i.e. it is the extent to which opinions on an issue are opposed, and the process by which this opposition increases over time)
Exaggeration	Article/Topic is represented as it is something as better or worse than it really is
Subjective	Text is written in a subjective way (opinion of the author) rather than journalistic.
Humoristic	The article uses humor as a writing style.

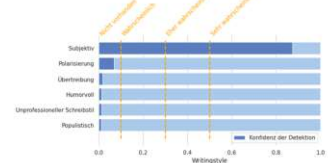
Writingstyle Detektions Klassifikation

Ergebnis: Überhebung: 0.16; Populistisch: 0.12; Polarisation: 0.06; Humorvoll: 0.10; Subjektiv: 0.73; Unprofessioneller Schreibstil: 0.03

In dem Text sind die folgenden Aspekte besonders herausgehoben:

Analyse Ergebnis

Text-Informationen



Relative Frequency of Detection

Text Content Assessment / Evaluation

Coverage of a wide range of semantic concepts

Name	Recognised contents	Language	Domain	Category Examples
Hate speech	Hatred against groups or individuals	Multi-ling	Social networks Discussion forums	Yes / No
Extremism	Extremist content	German	Social networks Article	Right-, Left-, Religious- or Single-Issue Extremism
Toxicity	Toxic, offensive content, comments, hateful language	German	Social networks	Yes / No
Factual assertions	Was it factually alleged?	Multi-ling	Social networks	Yes / No
Appealing contents	Appealing, positive, discussion-promoting language	German	Social networks Article	Yes / No
Sentimentality	Sentiment, feeling, emotion	German	Article	Positive, Negative
Report style	Report style of an article	German	Article	Conspiracy theory, clickbait
Writing style	Writing style of an article	German	Article	Polarise, exaggerate
Discrimination	Is a statement discriminatory?	German	Social networks	Ethnicity, social status
Relevance to criminal law	Is a statement criminal?	German	Social networks	Incitement, insult
Sexism	Various categories of sexism	English	Social networks	Misogyny, Sexual Violence

RAIDAR

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

Research project financed by the security-research program **KIRAS** by the Ministry of Finance in Austria

Timeline: **October 2021 – September 2023**

Topics: ***Hate Speech Detection, Extremism, Radicalization, Artificial Intelligence***

Coordination: AIT, Alexander Schindler (alexander.schindler@ait.ac.at)

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



CONTENT ASSESSMENT / EVALUATION

Questions:

- ⊗ How are these discussions conducted?
- ⊗ How much hate speech?
- ⊗ How much extremism?
- ⊗ What kind of extremism?
- ⊗ ...

- ⊗ Aggregation / Summarization of Machine Learning results
- ⊗ Trend / Time series Analysis
- ⊗ Graph Analytics

RAIDAR

Select a datasource:

Search for keywords:

Case sensitive

Limit the amount of docs, max: 100

Filter classification models:

- Criminal Paragraph
- Extrem
- Sweden Multiclass
- Toxicity Binary
- HateSpeechClassification
- EngagingComments
- Threat
- ExtremismBinary
- HateSpeechBinary
- Clan
- WumpDate
- ExtremismMultilabel
- CriminalParagraph
- ToxicityMulticlass
- SentimentMulticlass
- HateSpeechTarget
- Expression
- ReportingStyle
- HateSpeechMultilabel

Operator: OR

Probability threshold: 0.90

Filter images:

Select Symbolic Detection labels:
 HateSpeech

Operator: OR

Probability threshold: 0.98

Select a start date (From):
 Sep 11, 2018

Select an end date (To):
 Sep 11, 2018

Apply/Ok/cancel

Home Overview Datasets Images Trending

DerDritteWeg

This page gives an overview of each dataset. You can switch the dataset via "Select a datasource" on the left sidebar of the page.

The dataset contains a total of 12684 contents. This splits into 6055 images, 6029 text messages and 0 items of other types.

2020-01-11 - 2020-04-10

Category	Percentage
Images	47.74 %
Messages	52.26 %
Other	0 %

Message Overview

Category	Percentage
Hate Speech	55.83 %
Criminal Relevance	0.18 %
Toxicity	0.0 %

3763 out of 6829 processed messages (out of a total of 6829 messages) are classified as Hate Speech with a minimal probability of 30.0 %

12 out of 6829 processed messages (out of a total of 6829 messages) are classified as Criminal Relevance with a minimal probability of 30.0 %

0 out of 6829 processed messages (out of a total of 6829 messages) are classified as Toxicity with a minimal probability of 30.0 %

Extremism Multilabel Detector

For the multilabel detectors we show a lower threshold, since they are not predicting either/or the class but show whether there is a possibility of finding any of the labels. This means that one label does not exclude the other.

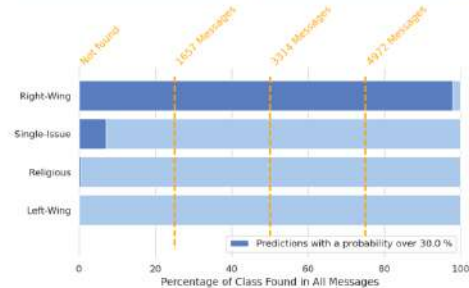


Image Overview

The dataset contains a total of 6055 images with 26 predicted as radical symbols, which is a total 0.43 %

Symbol	Percentage
Hakenkreuz	0.05 %
Doppel-Schlange	0.0 %
Sonnenrad	0.0 %
Wolfsangel	0.02 %
Keltenskreuz	0.36 %

3 out of 6055 processed images (out of a total of 6055 images) are predicted containing Hakenkreuz with a minimal probability of 30.0 %

0 out of 6055 processed images (out of a total of 6055 images) are predicted containing Doppel-Schlange with a minimal probability of 30.0 %

0 out of 6055 processed images (out of a total of 6055 images) are predicted containing Sonnenrad with a minimal probability of 30.0 %

1 out of 6055 processed images (out of a total of 6055 images) are predicted containing Wolfsangel with a minimal probability of 30.0 %

22 out of 6055 processed images (out of a total of 6055 images) are predicted containing Keltenskreuz with a minimal probability of 30.0 %

HYBRIS

Hybride Bedrohungs-Resilienz durch Interdisziplinäre Zusammenarbeit der Sicherheitsbehörden

Research project financed by the security-research program **KIRAS** by the Ministry of Finance in Austria

Timeline: **January 2023 – December 2024**

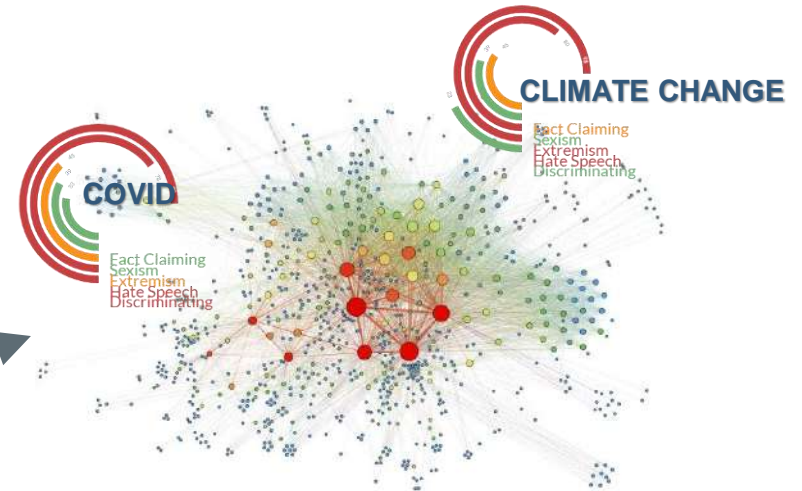
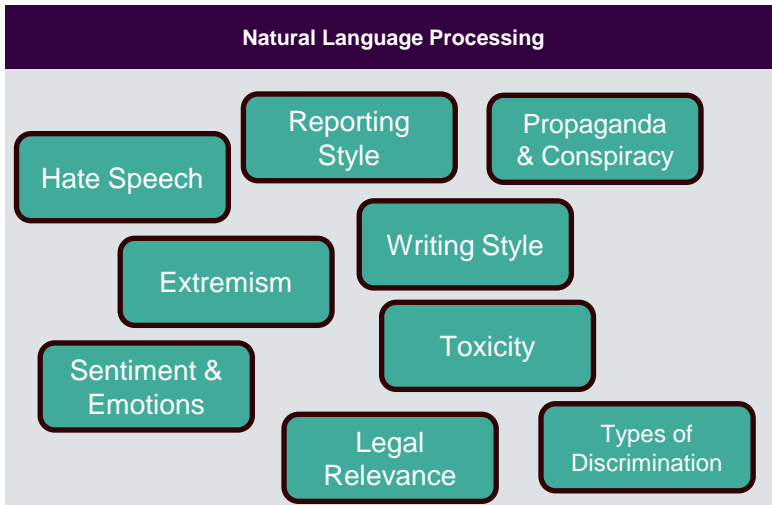
Topics: ***Disinformation Detection, Hybrid Threats, Artificial Intelligence***

Coordination: AIT, Alexander Schindler (alexander.schindler@ait.ac.at) & Mina Schütz



Information Nutrition Labels

- **Automatic Detection of Disinformation Using AI Models**
- Assessment of over 50 features.
- Detection of correlations and themes.



Situational Awareness

Critical Infrastructure Analysis Tool

Start date: 30.11.2023 | End date: 20.01.2024

Search keywords: [Enter any keywords...]

Sources with filter: [Enter any sources...]

Critical infrastructure

- critical infrastructure
 - Energy
 - district heating supply
 - gas supply
 - fuel and heating oil supply
 - electricity supply
 - Food
 - Finance and insurance
 - Health
 - Information Technology and Telecommunications
 - Media and culture
 - Municipal waste management
 - Government and administration
 - Transport and traffic
 - Water

Map is interactive (Click on a bar on the map to retrieve documents!)

Timeline: 2023-01 | 2023-03 | 2023-05 | 2023-07 | 2023-09 | 2023-11 | 2024-01

Documents

Title	Source	Weight
France's Orange withdraws from proposal to buy stake in British Telecom	altpress.fr	0.58
ComFutelle Enters Wind Weather Keep Ltd on Global Natural Gas Prices - LNG Recap - Natural Gas Intelligence	natgasintel.com	0.57
Cherries to Liberate Gas from ABC Resources under 13-Year Deal	rigzone.com	0.57
Space Global Awarded Space Services Contract by Lucara Space to Build and Operate Six Satellites for a Dedicated LEO Constellation	spacenews.com	0.57

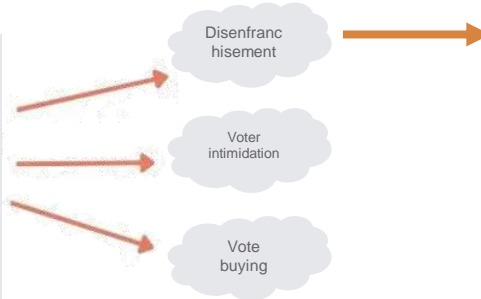
Threat Estimation & Narrative Detection

Taxonomie

Election fraud analysis

- ^ Electoral fraud
 - ^ Electorate manipulation
 - Artificial migration or party membership
 - Disenfranchisement
 - Division of opposition support
 - ^ Voter intimidation
 - Voter disinformation
 - Vote buying
- ^ Voting process and results
 - Misleading or confusing ballot papers
 - Ballot stuffing
 - Misrecording of votes
 - Misuse of proxy votes
 - Destruction or invalidation of ballots
 - Tampering with electronic voting systems
 - Voter impersonation
 - Postal ballot fraud

Nachrichten



Story-Card

Title: <LLM generated accumulated Title>

Num. Articles: 35

Key-Claims:

1. <LLM extracted key-claim>
2. <LLM extracted key-claim>
3. <LLM extracted key-claim>
4. <LLM extracted key-claim>
5. <LLM extracted key-claim>

Indicators:

- Incitement of Violence
- Economic Implications
- Medical Implications



Narratives:

- Central Narrative:
- Further Narrative

Predjudice:

- Antisemitism
- Anti-Feminism





Thank you!

Mina Schütz

Data Science & Artificial Intelligence
Center for Digital Safety & Security

mina.schuetz@ait.ac.at | www.ait.ac.at

STARLIGHT



This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101021797