AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Fake News Detection

and detecting other Toxic content on the Web

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

AIT Austrian Institute of Technology GmbH
Giefinggasse 4 | 1210 Vienna | Austria
T +43 50550-2902 | M +43 664 8251454
alexander.schindler@ait.ac.at | www.ait.ac.at

Alexander Schindler

Dipl.-Ing. Dr. techn. / Bakk.techn.

AUSTRIAN INSTITUTE OF TECHNOLOGY

Senior Scientist / Thematic Coordinator Data Science / Deputy Head of Competence Unit

- Multi-Modal Machine Learning, Applied Artificial Intelligence
- Team Lead: Audio Processing, NLP

External Lecturer @ TU-Wien

- Data Science, Information Retrieval, Intelligent Audio and Music Analysis, Student Supervision

External Lecturer @ FH St. Pölten (until 2023)

- Selected Topics of Machine Learning and AI

Research Interests

- Audio / Music Analysis, Audio-Visual Analysis, Machine Learning / Deep Learning, Artificial Intelligence

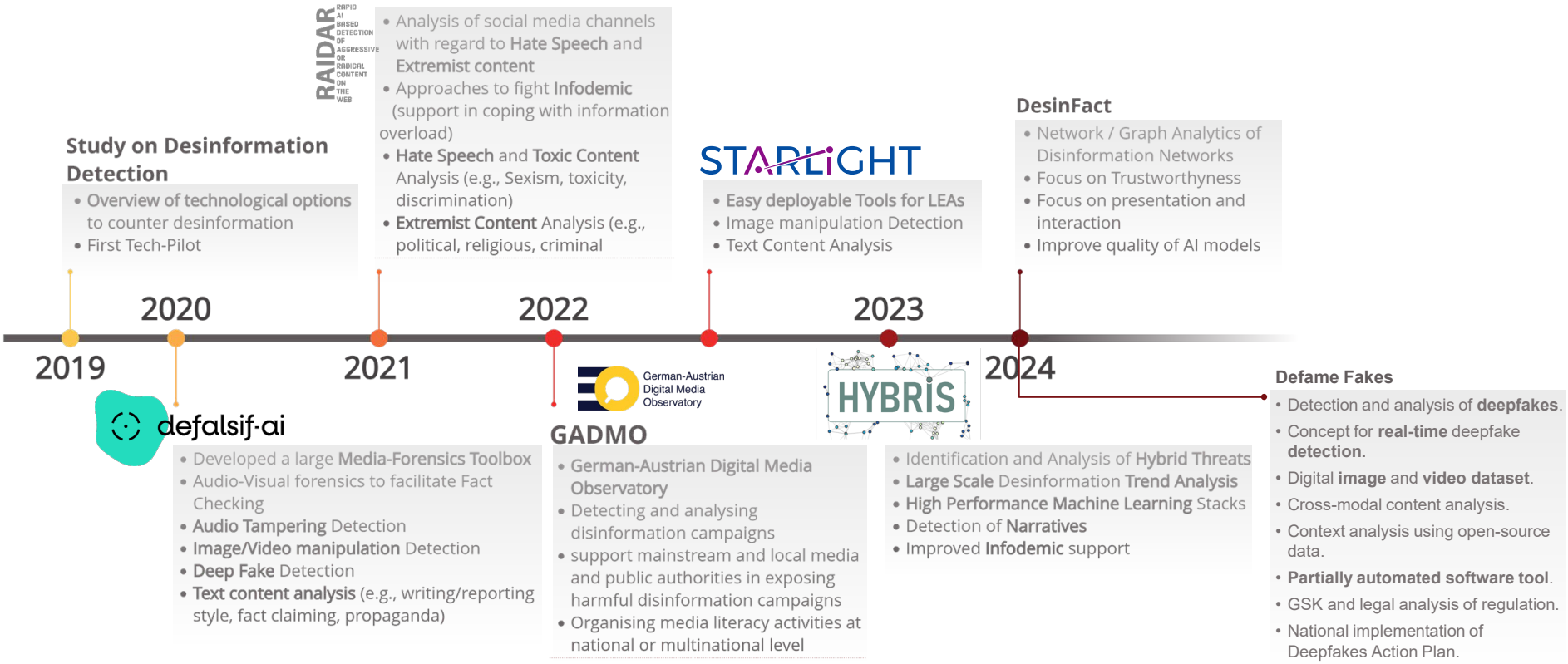
Event Organization

- Vienna Deep Learning Meetup
- AI-Summit 2017, Ethics & Bias in AI 2018, WeAreDevelopers AI Congress 2018 (Partner)
- Tutorials on Deep Learning (ML-Prague 2018, ISMIR 2018)
- Int. Workshop on Music Speech and Mind (2019 - 2021)
- Int. Workshop on Disinformation and Toxic Content Detection (DiTox 2023)

Disinformation Detection Research @ AIT

Line of Research Projects

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY



Disinformation Detection Research @ AIT

Research Team

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Dr. Alexander Schindler
Project Lead

DI. Martin Boyer
Project Lead

NLP

Mina Schütz, BSc. MSc.
NLP Models
knowledge graphs
Visual Analytics

Medina Andresel, MSc.
NLP Models
Logic Reasoning
Knowledge Graphs

Daria Liakhovets, BSc. MSc.
NLP
System Integration
Project Managmenet

DI. Simon Ott
NLP
Knowledge Graphs

Computer Vision

Dr. David Fischinger
Computer Vision Models

Silvia Poletti, MSc.
Computer Vision Models

High Performance Computing

Mihai Bartha
System Integration
High Performance Computing
Software Engineering

Audio

Lam Pham, PhD.
Audio Analytics
Machine Learning

AIT AI Ethics Lab

PD Dr. Peter Biegelbauer
Head of AIT AI Ethics Lab

Mag. Pia Weinlinger
Ethical assessment

Rodrigo Conde-Jimenez, MSc.
Ethical assessment

Disinformation Detection Research @ AIT
Cooperation Partner

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

MINISTERIAL COOPERATION

- Federal Chancellery
- Federal Ministry Republic of Austria Justice
- Federal Ministry Republic of Austria European and International Affairs
- Federal Ministry Republic of Austria Defence
- Federal Ministry Republic of Austria Interior

INSTITUTIONAL COOPERATION

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

RESEARCH AND INDUSTRY PARTNERSHIPS

LIquA Linzer Institut für qualitative Analysen · 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 Athens Technology Center · technische universität dortmund

FUNDING PROGRAMS

FFG Promoting Innovation. · KIRAS Security Research · European Commission | Horizon 2020 European Union funding for Research & Innovation

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

How Fake News & Disinformation work

Overview

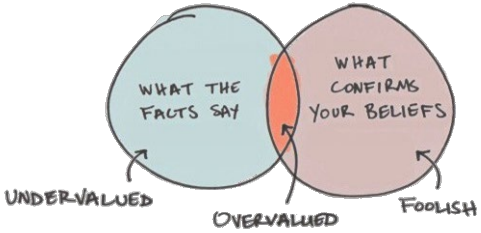
Desinformation

- Disinformation is **not an invention of the 21st century**, but…
- the **rapid growth of social media** and the **Internet has made it easier for false information to spread** quickly and widely
- The **rise of generative AI**, like deepfakes and ChatGPT, **blurs the lines between real and fake**, making discernment increasingly challenging.
- A **climate of political polarization** and **eroding trust in traditional media** has fostered an environment conducive to the flourishing of disinformation.
- **Disinformation campaigns** are frequently executed by influential entities, including governments and political organizations, with a clear objective: manipulate public opinion and undermine democratic processes.
- **Confirmation bias** plays a pivotal role, as individuals tend to propagate or accept information that aligns with their existing beliefs.

Ideological Priors

or

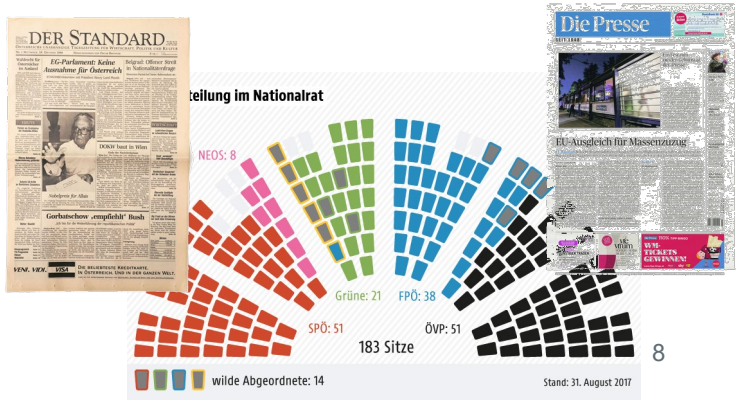
Why do we fall for Fake News?



Confirmation Bias

- **Naive Realism**
  - Believe information that is aligned with your views
- **Confirmation Bias**
  - Seek information that confirms your existing views
- **Normative Influence Theory**
  - Consume/Share socially safe options ➔ for social acceptance, affirmation

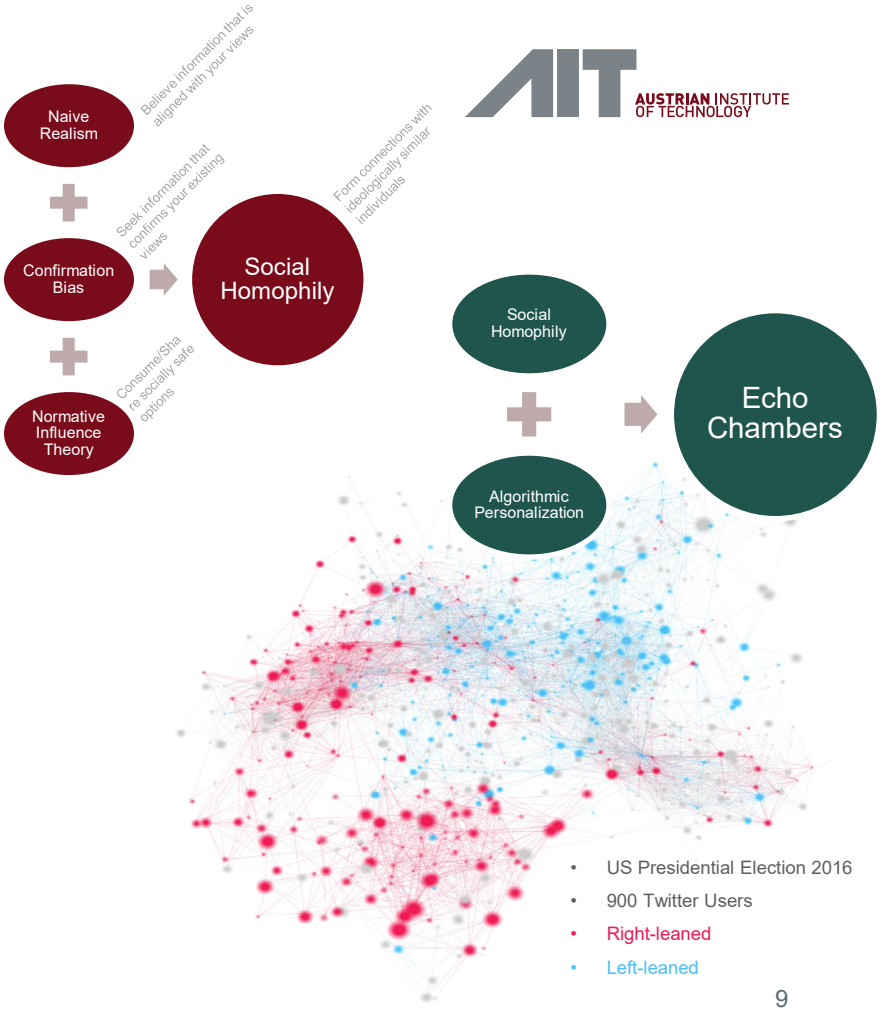
➔ Individual level of Fake News



Nature / Characteristics

- **Social Level**

- **Echo Chambers / Filter Bubbles**
  - **Social homophily**
    - Form connections with ideologically similar individuals
  - **Algorithmic personalization**
    - Read content
    - Follow / befriend persons
  - ➢ Consequences
    - ➢ Less exposure to conflicting viewpoints
    - ➢ Isolation in own filter bubble
    - ➢ Improve survival / spread of fake news



Gillani, Nabeel, et al. "Me, my echo chamber, and I: Introspection on social media polarization." Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, 2018.

The Role of Social Media

Amplification

- Algorithms that incentivize content sharing contribute to the proliferation of false information, increasing its visibility and reach.

Rapid Dissemination

- False information spreads swiftly and widely, facilitated by digital platforms.

Anonymity

- The ability to disseminate false information anonymously hinders accountability for its sources.

Viral Sensationalism

- False information with sensational or attention-grabbing elements tends to go viral more easily.

Filter Bubbles

- Users often encounter information that aligns with their existing beliefs, creating challenges when attempting to confront false information within their information bubbles.

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

What Is Fake News, Disinformation, Misinformation?

Overview & Definitions

Disinformation Concepts

Misinformation	Disinformation	Propaganda	Fabrications
• Unintentional • Mix of facts & fiction • Includes satire, fabrications, false context	• Intentional • User posts it regardless of being false • Deceiving	• Intentionally misleading • Biased • Influencing opinions of users	• Synonym to Fake News • looks legitimate • Yellow press / tabloids

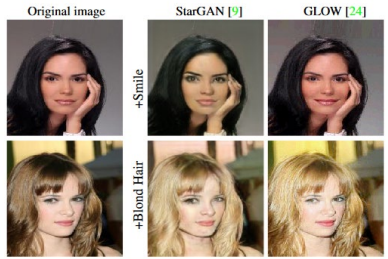
Satire	Hoax	Rumor	Clickbait
• Unintentional • Based on exaggeration and humor • Entertainment	• Fabrications • Deceive audiences • Large hoaxes hard to detect	• Not verified • Can be true • After new events	• Revenue through clicks • Eye-catching headlines • Unstructured

Intention / Facticity

- **Fabricated content**
  - *Completely false*
- **Misleading content**
  - *Misleading use / framing of issue*
- **Imposter content**
  - *Genuine source impersonated with false sources*
- **Manipulated content**
  - *For deception (e.g. images)*
- **False connection**
  - *Headlines, visuals do not support content*
- **False context**
  - *Genuine content shared with false context information*

Fabricated Content

False Context



Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.

Manipulated Content

Imposter Content

How Your Audience Will Believe Anything: The Psychology Behind the Fake News https://www.click.co.uk/blog/how-your-audience-will-believe-anything-the-psychology-behind-the-fake-news-bas-van-den-belds-benchmark-2018-talk-review/

Generative AI

- **Voice Style transfer**
- **Image generation**

https://www.youtube.com/watch?v=SIWKf2sZfok

DEFINITION BY STUDIES

The European Commission defines disinformation as the creation, presentation and dissemination of verifiably false or misleading information for the purposes of economic gain or intentionally deceiving the public, and which may cause public harm. Such harm may include undermining democratic processes or threats to public goods such as health, the environment and security.

As opposed to illegal content (which includes hate speech, terrorist content or child sexual abuse material), disinformation covers content that is legal. It therefore intersects with fundamental core European Union (EU) values of freedom of expression and the press. Under the Commission's definition, disinformation does not include misleading advertising, reporting errors, satire and parody, or clearly identified partisan news and commentary.

Threat scenarios / Impact

Misleading the Public

- The dissemination of inaccurate information can influence people's decisions and beliefs.

Undermining Democracy

- False information can be weaponized to manipulate elections and erode confidence in democratic institutions.

Aggravation of conflicts

- Disinformation has the potential to stoke social and political tensions, escalating conflicts and violence.

Harming Reputations

- Rumors and unfounded claims are often weaponized to tarnish the reputations of individuals and organizations.

Endangering Public Health

- Conspiracy theories and unfounded medical advice can pose serious risks to public health by promoting unscientific beliefs about causes, symptoms, and treatments.

Manipulation of Slovakian Elections

Slovakia's Election Deepfakes Show AI Is a Danger to Democracy

Fact-checkers scrambled to deal with faked audio recordings released days before a tight election, in a warning for other countries with looming votes.

Progressive Slovakia party leader Michal Simecka. PHOTOGRAPH: ZUZANA GOGOVA/GETTY IMAGES

- Progressive party leader leading in pre-election polls
- Recording of allegidly „attempted voting fraud“ with voices from party leader and journalist
- Published within 48 hours moratorium ahead of polls – no possibility to clarify
- Progressive party lost election

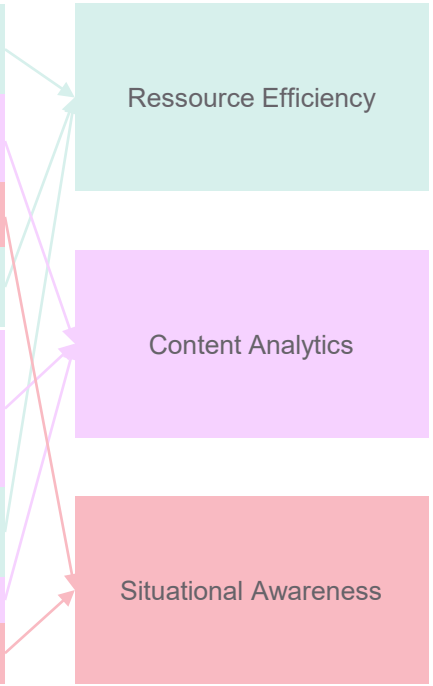
CHALLENGES OF THE INFORMATION AGE

- **Amount** of information.
- Rate of **new information being produced**
- **Continuous news culture (speed over quality)**
- Ease of **duplication and transmission**
- **More channels of incoming information**
- Ever-increasing amounts of **historical information** to dig through
- **Contradictions and inaccuracies in available** information
- A low **signal-to-noise ratio**
- Information **unrelated or lacking context**
- Speed/low cost at which information **can created, reproduced, and delivered.**
- Ease with **which information can be altered.**
- **Weaponization of information.**

CHALLENGES OF THE INFORMATION AGE

Requirements for a Detection System

- **Amount** of information.
- Rate of **new information being produced**
- **Continuous news culture (speed over quality)**
- Ease of **duplication and transmission**
- **More channels of incoming information**
- Ever-increasing amounts of **historical information** to dig through
- **Contradictions and inaccuracies in available** information
- A low **signal-to-noise ratio**
- Information **unrelated or lacking context**
- Speed/low cost at which information **can created, reproduced, and delivered.**
- Ease with **which information can be altered.**
- **Weaponization of information.**

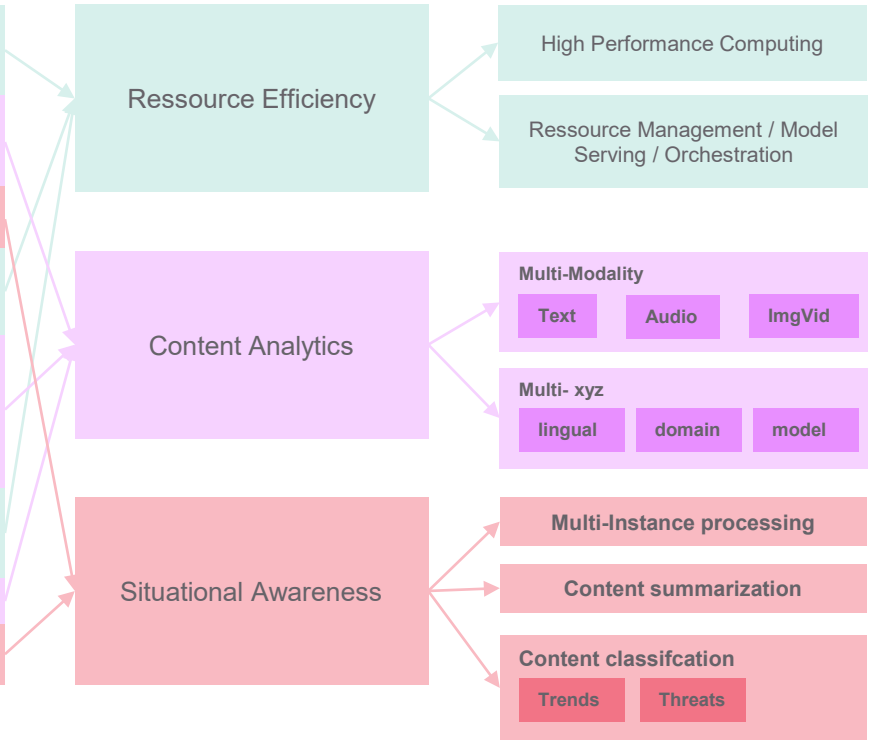


CHALLENGES OF THE INFORMATION AGE

AUSTRIAN INSTITUTE OF TECHNOLOGY

Requirements for a Detection System

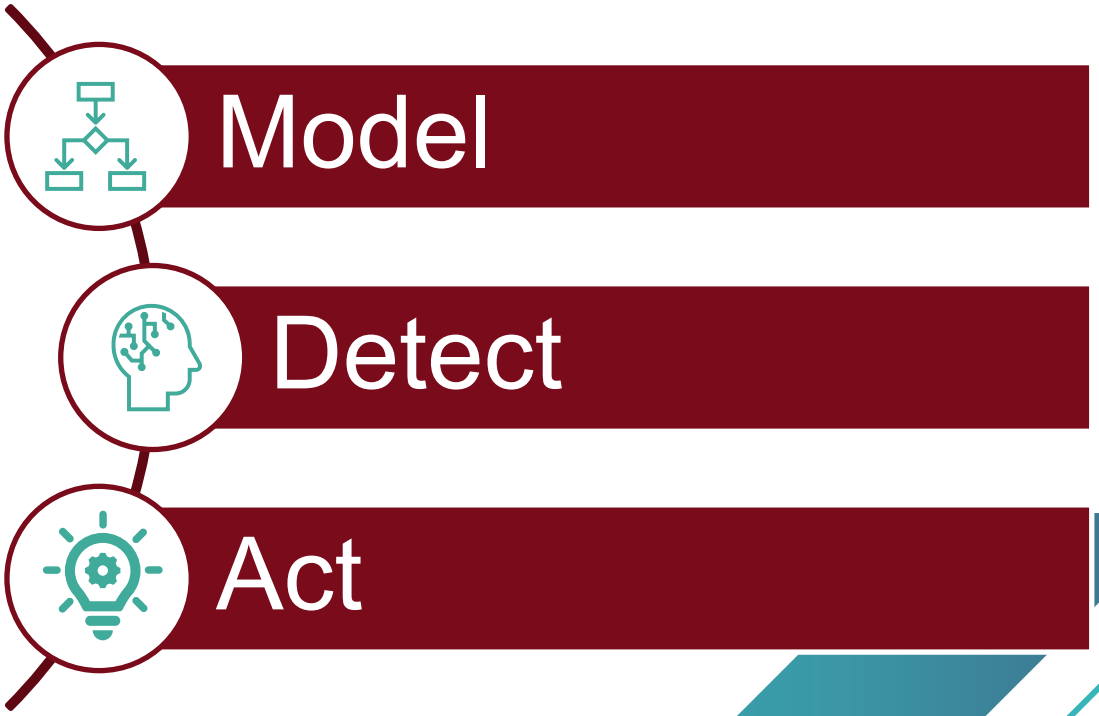
- **Amount** of information.
- Rate of **new information being produced**
- **Continuous news culture (speed over quality)**
- Ease of **duplication and transmission**
- **More channels of incoming information**
- Ever-increasing amounts of **historical information** to dig through
- **Contradictions and inaccuracies in available** information
- A low **signal-to-noise ratio**
- Information **unrelated or lacking context**
- Speed/low cost at which information **can created, reproduced, and delivered.**
- Ease with **which information can be altered.**
- **Weaponization of information.**



Countering Fake News & Disinformation

How can we… ?

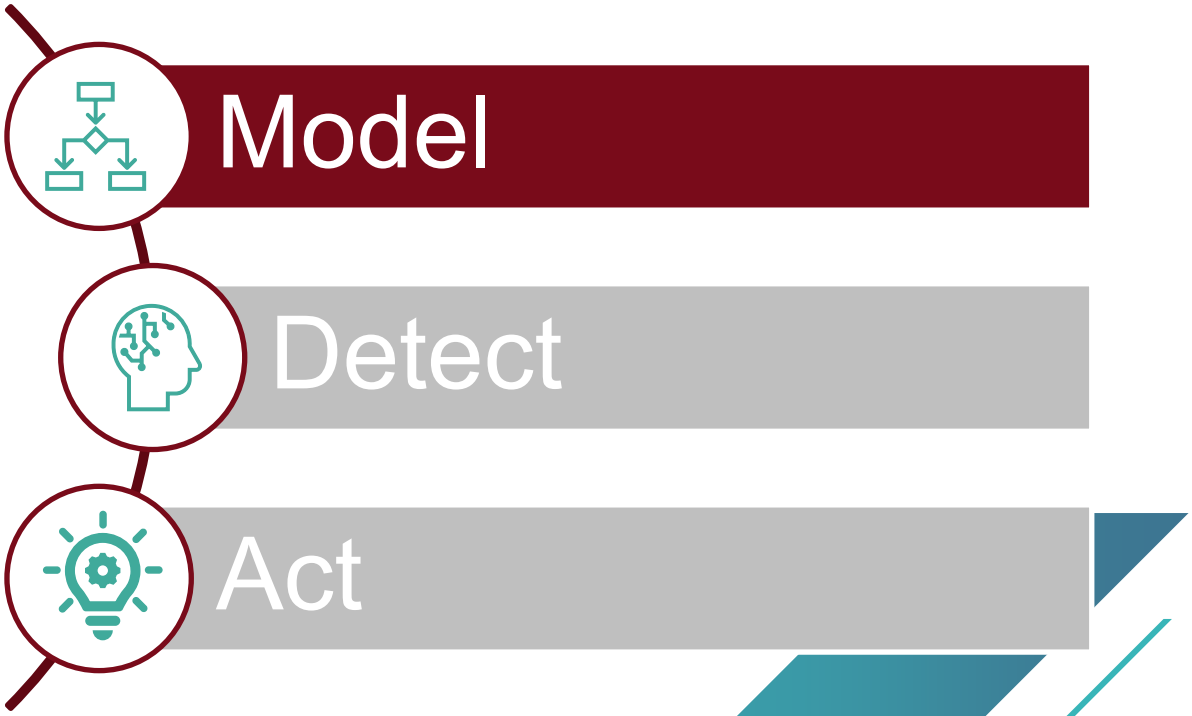
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY



Countering Fake News & Disinformation

How can we… ?

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Modelling (dis)information

Decomposing Information into measurable signals of credibility

The CCC Model

Contributor

- What can we find about the source of information?

Content

- Does the posted content look reliable?

Context

- Does everything contextualise together?

Contributor

Reputation

- what do people think of this source?

History

- what is the past activity of this source?

Presence

- where does this source exist?

Influence

- what happens because of this source?

Popularity

- who follows this source?

Contributor

Reputation
- what do people think of this source?

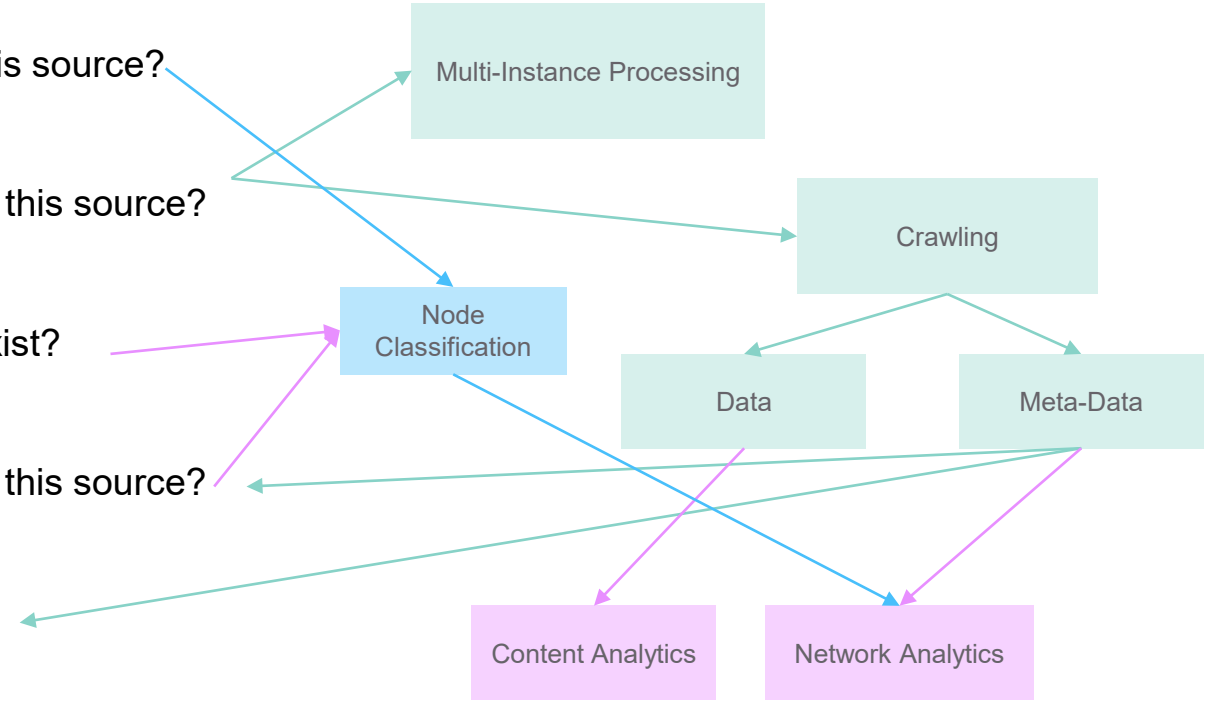
History
- what is the past activity of this source?

Presence
- where does this source exist?

Influence
- what happens because of this source?

Popularity
- who follows this source?

Requirements for a Detection System



Content

Quality

- what is the text/visual/audio style like?

Popularity

- what is the social interaction with it?

Authenticity

- has it been manipulated/synthetically generated?

History

- can it be found in past publications?

Reputation

- how is it referenced by others?

Content

Quality
- what is the text/visual/audio style like?

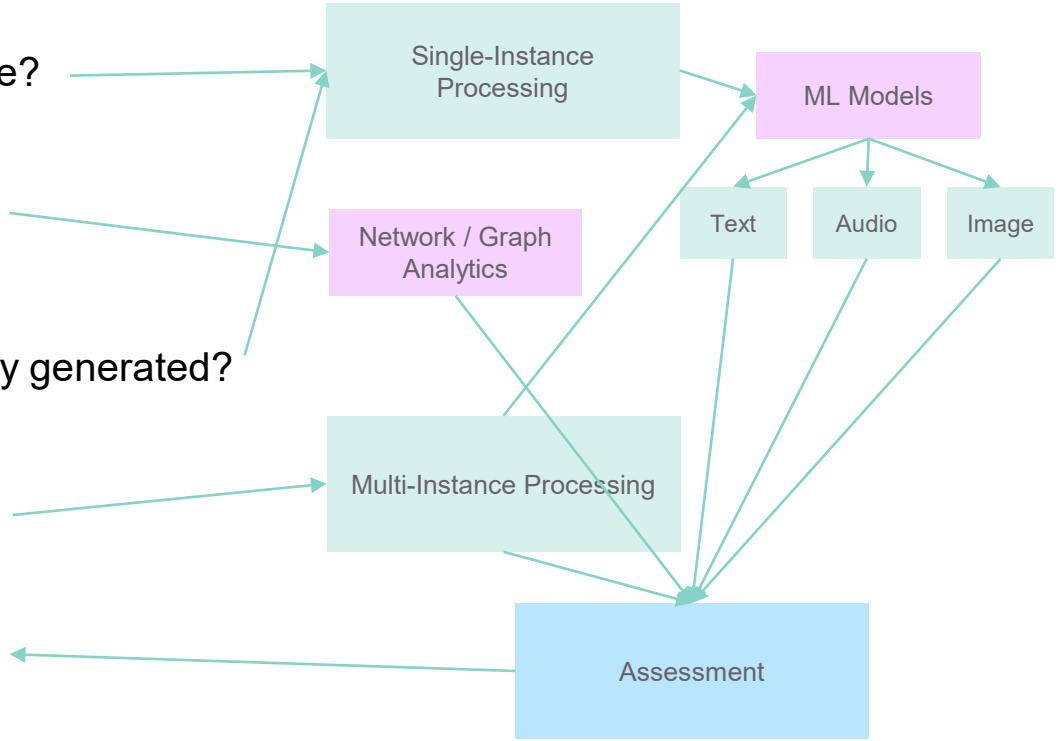
Popularity
- what is the social interaction with it?

Authenticity
- has it been manipulated/synthetically generated?

History
- can it be found in past publications?

Reputation
- how is it referenced by others?

Requirements for a Detection System



Context

Cross-check

- are there any similar reports?

Diversity

- are there multiple coherent reports?

Provenance

- how has this travelled through time?

Influence

- what happens because of this?

Proximity

- do source locations relate to events?

Context

Requirements for a Detection System

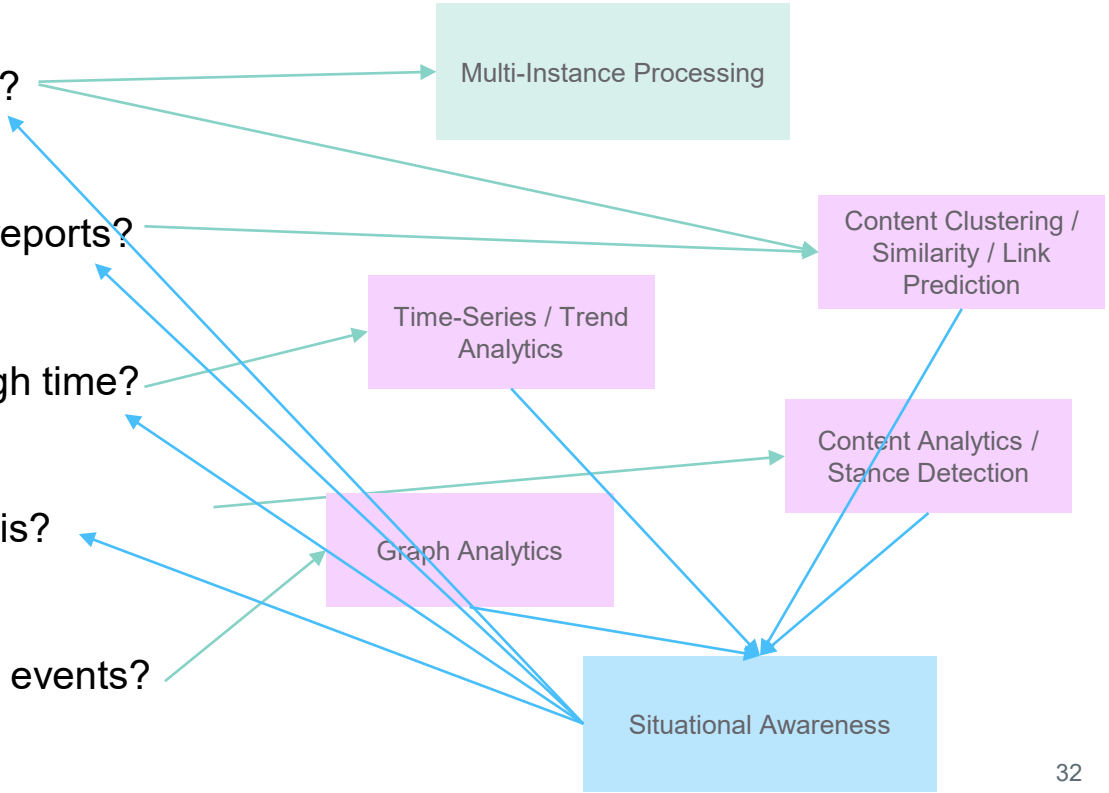
Cross-check
- are there any similar reports?

Diversity
- are there multiple coherent reports?

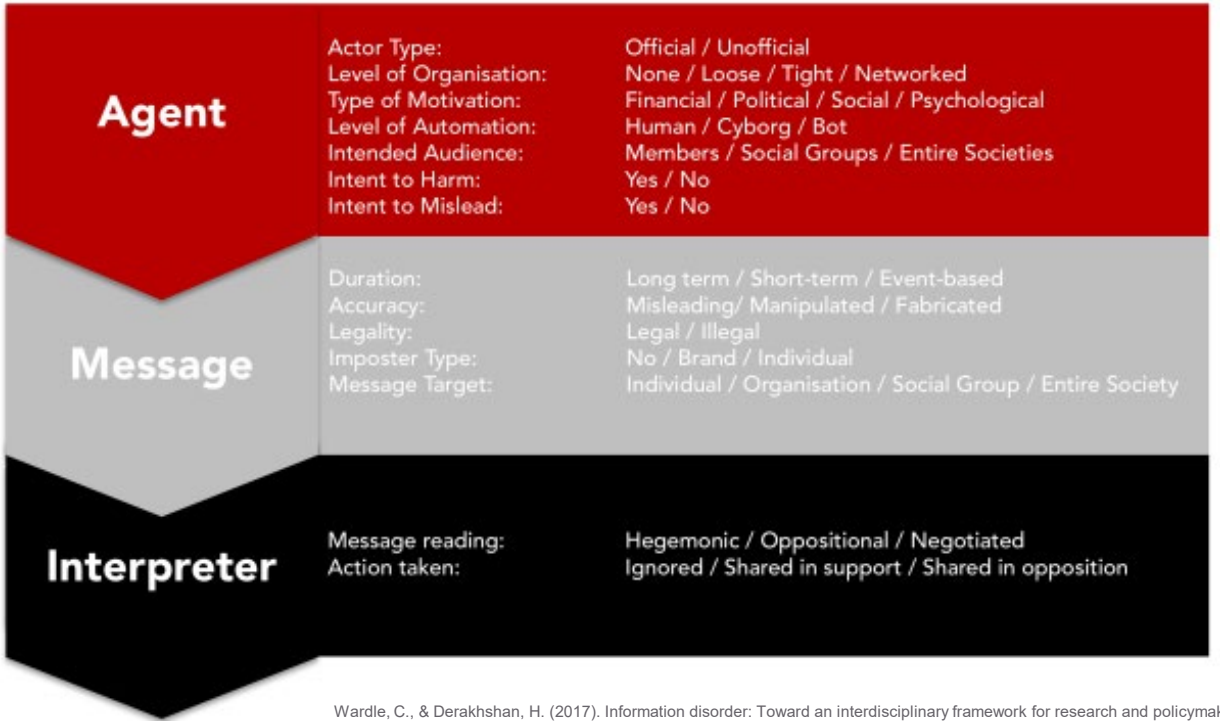
Provenance
- how has this travelled through time?

Influence
- what happens because of this?

Proximity
- do source locations relate to events?



The AMI Model

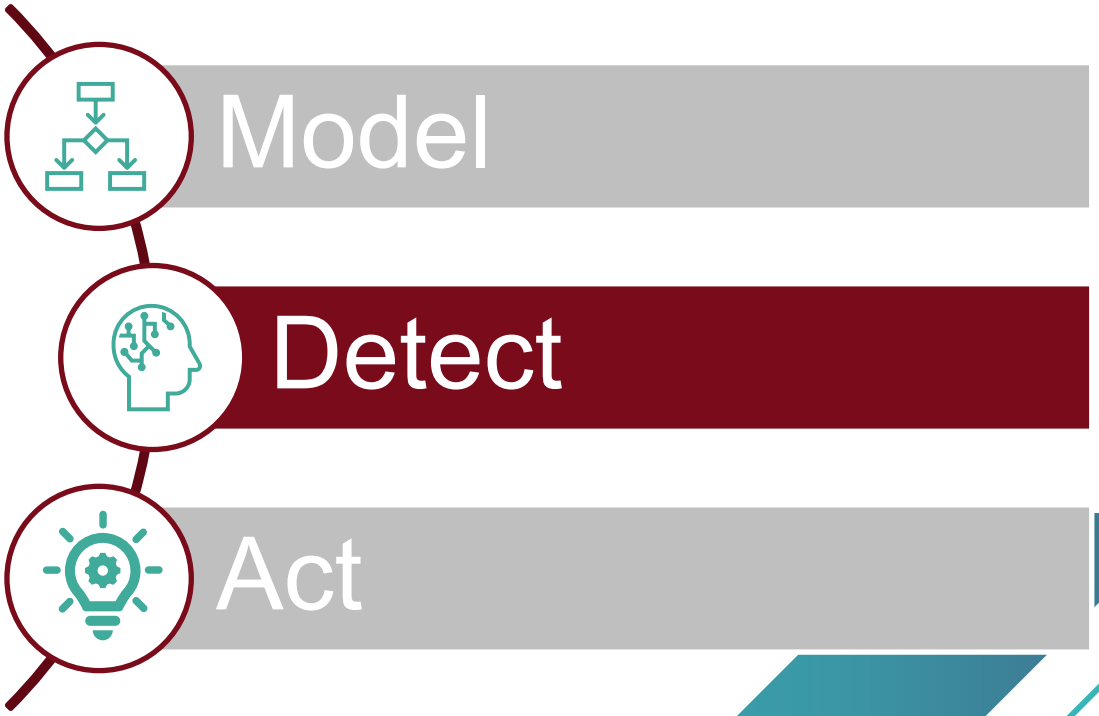


Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Countering Fake News & Disinformation

How can we… ?

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY



Detecting (dis)information Signals

Detect Signals and measure credibility

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

VISUAL COMPUTING ASSISTANCE

For Fact Checkers

Visual Signals

AUSTRIAN INSTITUTE OF TECHNOLOGY

Bildrückwärtssuche

Ergebnis ausblenden

Ergebnis: verdächtig

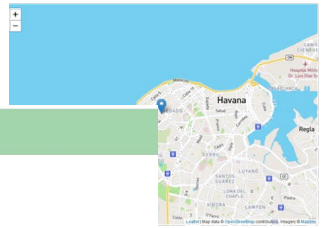
Verdächtig: Zeitunterschied ist 1494 Tage. Da dieses Bild schon 1494 Tage vor dem von Ihnen eingetragenen Datum (2022-09-25) online zu finden war, kann es sich um kein aktuelles Foto handeln.

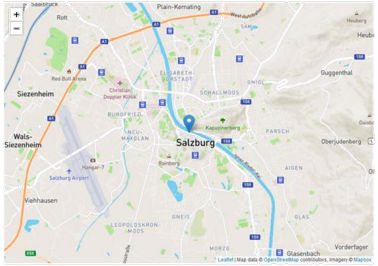
Analyse Ergebnis

Bild-Informationen

Alternative Instanzen

- Place de la Bourse - Wikipedia
- Place de la Bourse in Bordeaux - Expedia.at
- Die 5 schönsten Orte in Bordeaux
- Urlaub in Bordeaux | Bordeaux Magazin
- Aftermovie 2022
- 5 Roadtrips in Frankreich: Von der Bretagne bis an die Côte d ...





Fake-Face-Detektor

Ergebnis: fake

Das Bild wurde als künstlich klassifiziert.
Die roten Flächen weisen auf verdächtige Bildtexturen hin.

Analyse Ergebnis

Input Image

predicted: FAKE

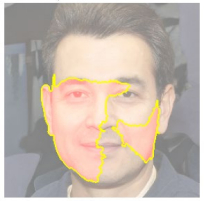


Bild-Manipulations-Detektor

Ergebnis: manipuliert

Das Bild wurde als manipuliert klassifiziert.

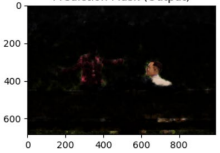
Analyse Ergebnis

Bild-Informationen

Decoded /defalsifnfs/data/airflow/input/Merkel.png of size (684, 992, 3) for 3.13 seconds

Image (Input)

Prediction Mask (Output)



Meta-Data Analysis

- **Task:**
  - Detect irregularities in image metadata
  - Check for consistency, plausibility, etc.

- **Method:**
  - Check for meta-data entries/traces left by software
    - E.g., Photoshop, GIMP, etc.
  - Check / Cross-Check
    - Correct dates, times (e.g., creation, modification)
      - file-, EXIF-, etc. meta-data
    - GPS coordinates

PNG:PixelUnits	1
PNG:ModifyDate	2013:03:10 19:48:43
PNG:Comment	Created with GIMP

Metadaten-Extraktion

Ergebnis: Metadaten extrahiert

Auf Basis der Metadaten ist die Quelle: unbekannt

Analyse Ergebnis

Meta-Informationen

Eigenschaft	Wert
EXIF:Make	Apple
EXIF:Model	iPhone 6
EXIF:Software	12.5.5
EXIF:ModifyDate	2021:09:27 11:22:24
EXIF:DateTimeOriginal	2021:09:27 11:22:24
EXIF:GPSLatitude	48.3523972222222
EXIF:GPSLongitude	16.3283916666667
EXIF:GPSAltitude	167.4697419
EXIF:GPSTimeStamp	09:22:23.23
EXIF:GPSDateStamp	2021:09:27
SourceFile	/defalsifnfs/data/airflow/input/ImageWithGPSData2.jpg
ExifTool:ExifToolVersion	12.4

Meta-Data / Claim Analysis

- **Task:**
  - Cross-Check claims
    - E.g., about location, weather, demography
- **Method:**
  - Check online/historic weather information
  - Sun position
  - OpenStreetMap, Google Maps, Google Street View, Google Earth

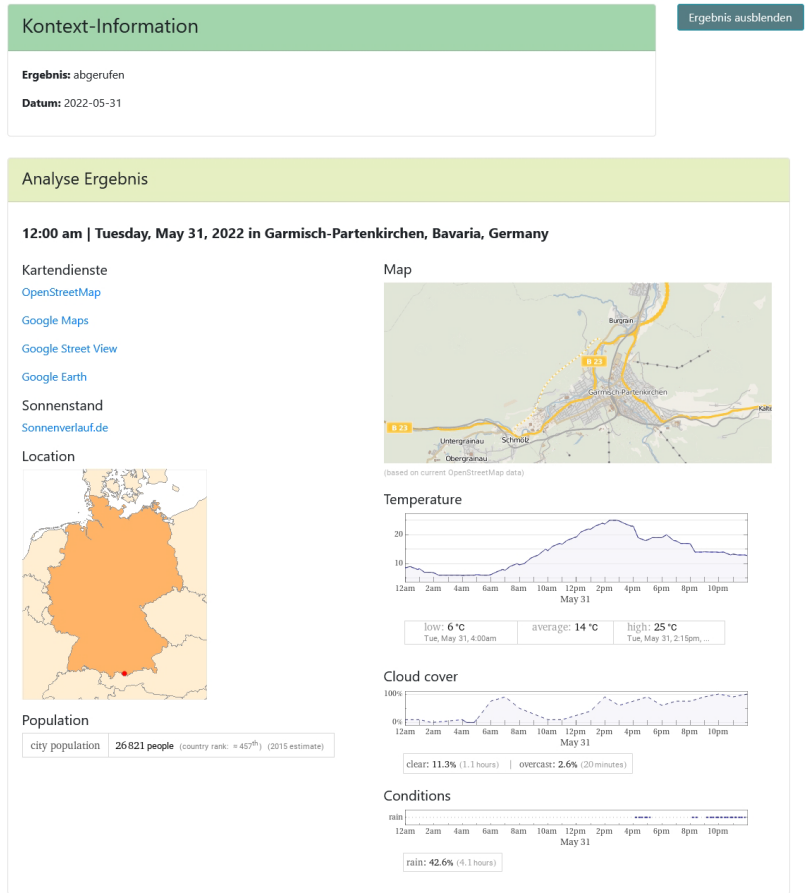
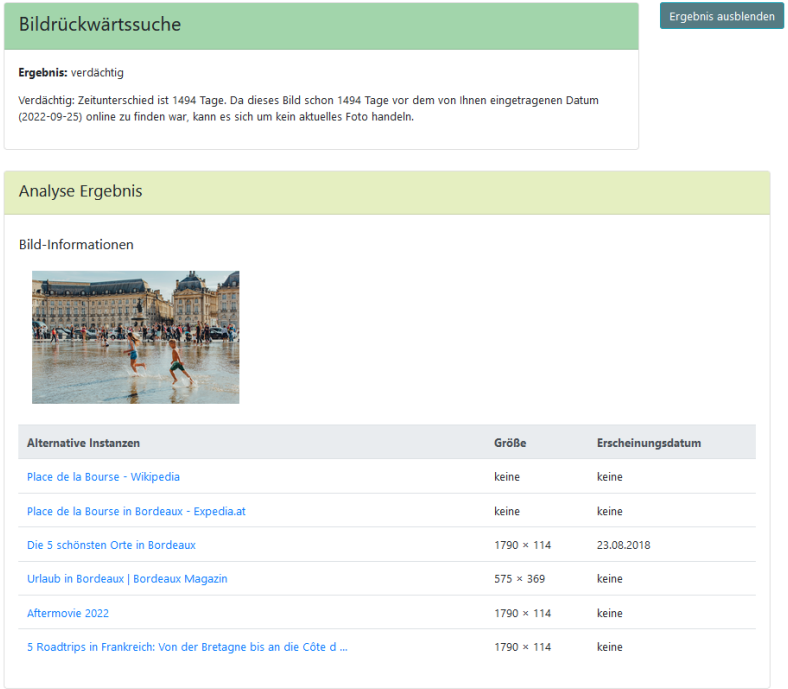


Image Reverse Search

- **Task:**
  - Check if an image
    - is original or from another website
    - has originally been published in a different context
  - Notification of conspicuous time differences to current date or date entered by user (False context)
  - Often useful for determining location

- **Threat Scenarios:**
  - Images are reused/repurposed in a different context

- **Method:**
  - Google, Bing, etc. reverse image search



Fake Face Detection

- **Task:**
  - Detect generated images
- **Threat Scenarios:**
  - Fake user-profiles, author images, witnesses, experts
- **Method:**
  - Dataset created with 125,000 images from various sources:
    - 47K Real: FFHQ
    - 78K Fake: PGGAN, Stylegan, Stylegan 2
  - Training of Deep Neural Network (binary decision fake/real)
  - Classification of artificially generated faces on benchmark datasets: 95%-99.8% correctness
  - Visual Explainability

https://thispersondoesnotexist.com/

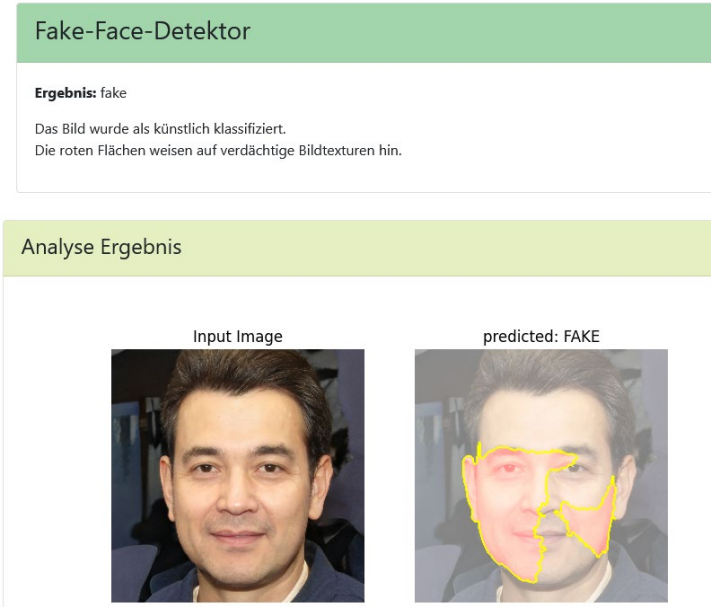


Image Forgery Detection

AUSTRIAN INSTITUTE OF TECHNOLOGY

- **Task:**
  - Detect image tampering
    - Copy-Move (C)
    - Splicing (S)
    - Removal (R)
    - Enhancement (E)

Manipulation-Type	C	S	R	E
Resample	×	×	–	–
Flip	×	×	–	–
Rotate	×	×	–	–
Blur	–	–	–	×
Contrast	–	–	–	×
Noise	–	–	–	×
Brightness	–	–	–	×
JPEG-Compression	–	–	–	×

- **Threat Scenarios:**
  - All disinformation types and categories

Image Forgery Detection

AUSTRIAN INSTITUTE OF TECHNOLOGY

- **Method:**
  - 11 data sets generated on the basis of MS COCO
  - >2,000,000 images (224x224)

- Image Masks:

Shape of Mask
Triangle
Rounded Rectangle
Ellipse
Polygon with 5 vertices
Ellipse + Polygon with 4 vertices
Superpixel Segmentation
Person Segmentation

- Generated neural network extracts image features and detects local anomalies.

- Result indicates which areas in the image have been potentially manipulated.

Example Training Images

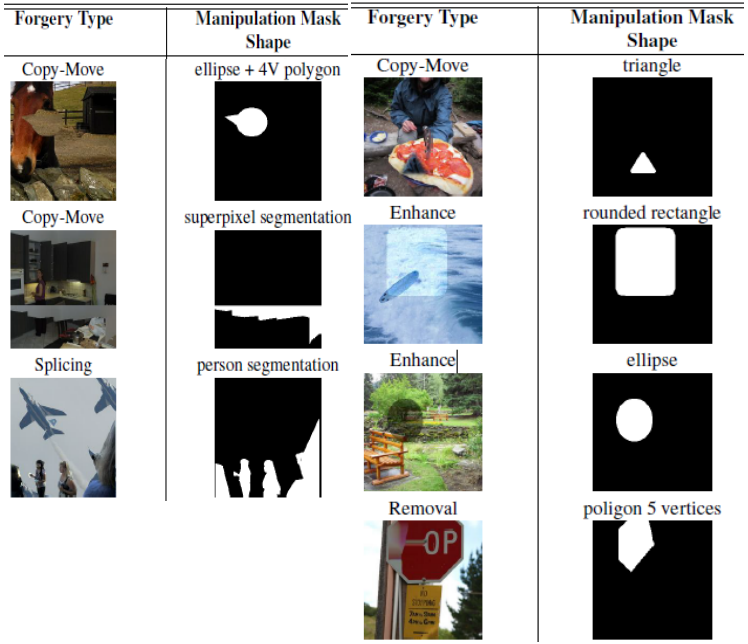
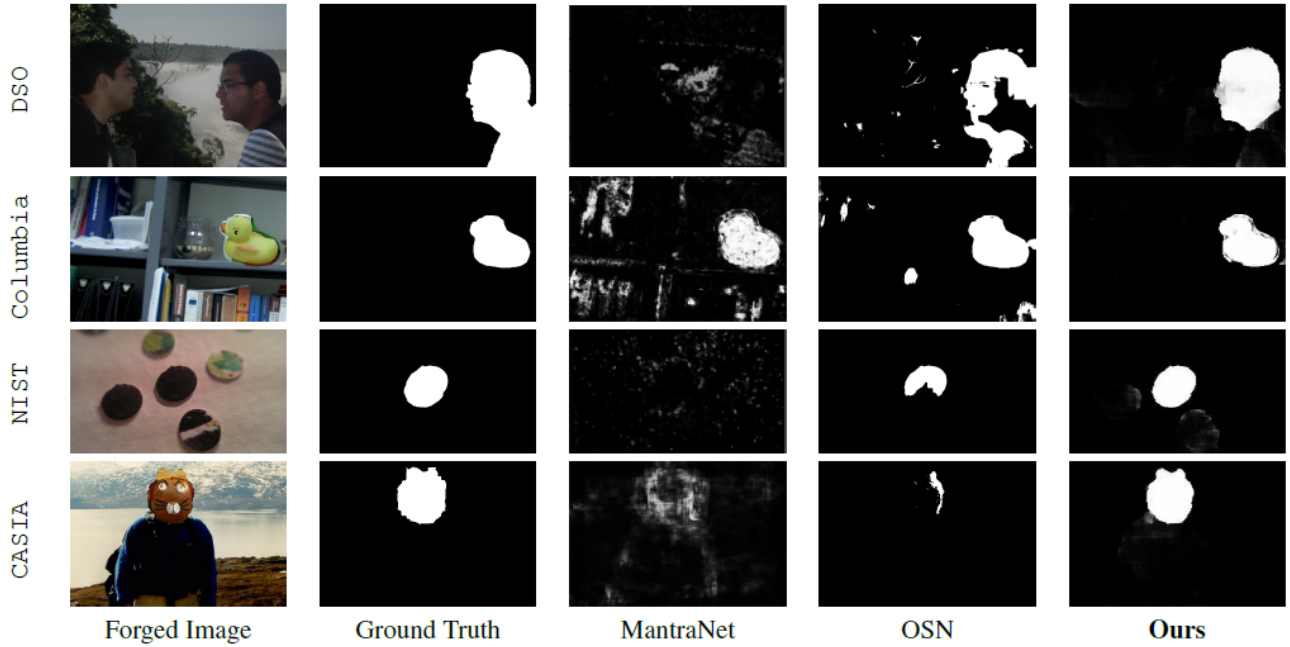


Image Forgery Detection

AUSTRIAN INSTITUTE OF TECHNOLOGY

Manipulation detection - evaluation: examples

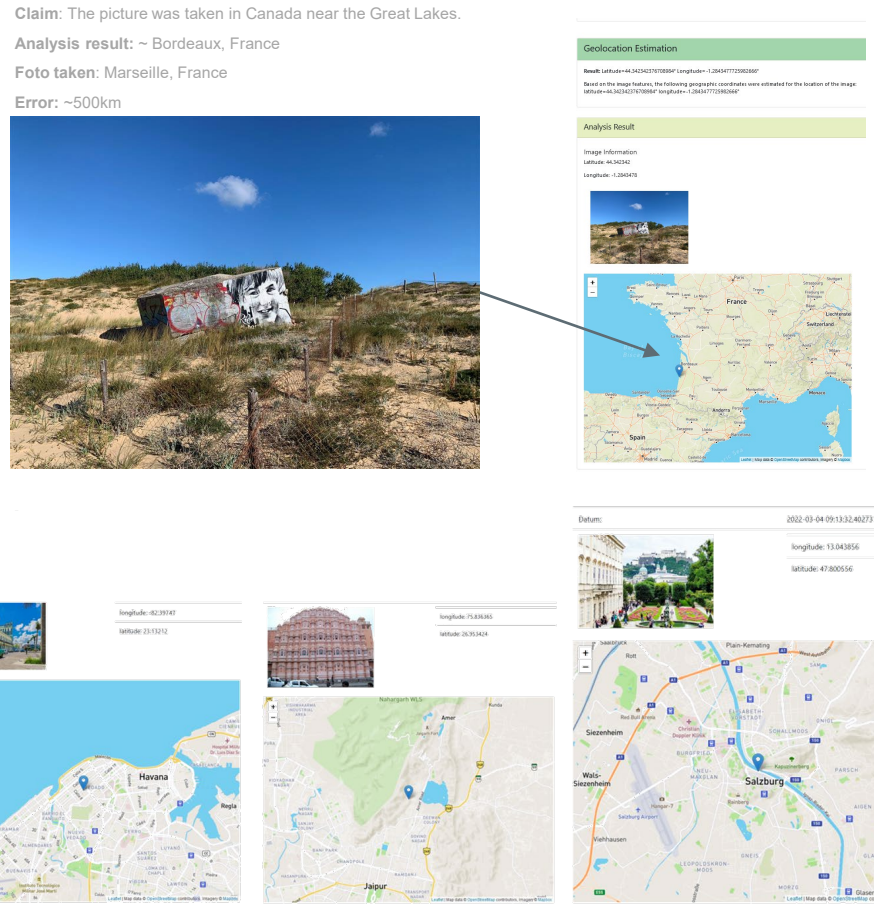


- Fischinger, D. and Boyer, M. (2023). DF-Net: **The digital forensics network for image forgery detection**. Irish Machine Vision and Image Processing conference.
- Fischinger, D. and Boyer, M. (2023). DF2023: **The digital forensics 2023 dataset for image forgery detection**. Irish Machine Vision and Image Processing conference.

Models	AUC of Test Datasets				
	DSO [3]	Columbia [6]	NIST [15]	CASIA [4]	Average
ForSim [14]	**.796**	.731	.642	.554	.681
DFCN [29]	.724	.789	.778	.654	.736
ManTra-Net [24]	.795	.747	.634	.776	.738
OSN [23]	.723	.815	.686	.751	.744
SE_UN (ours)	.732	**.827**	**.780**	**.851**	**.797**

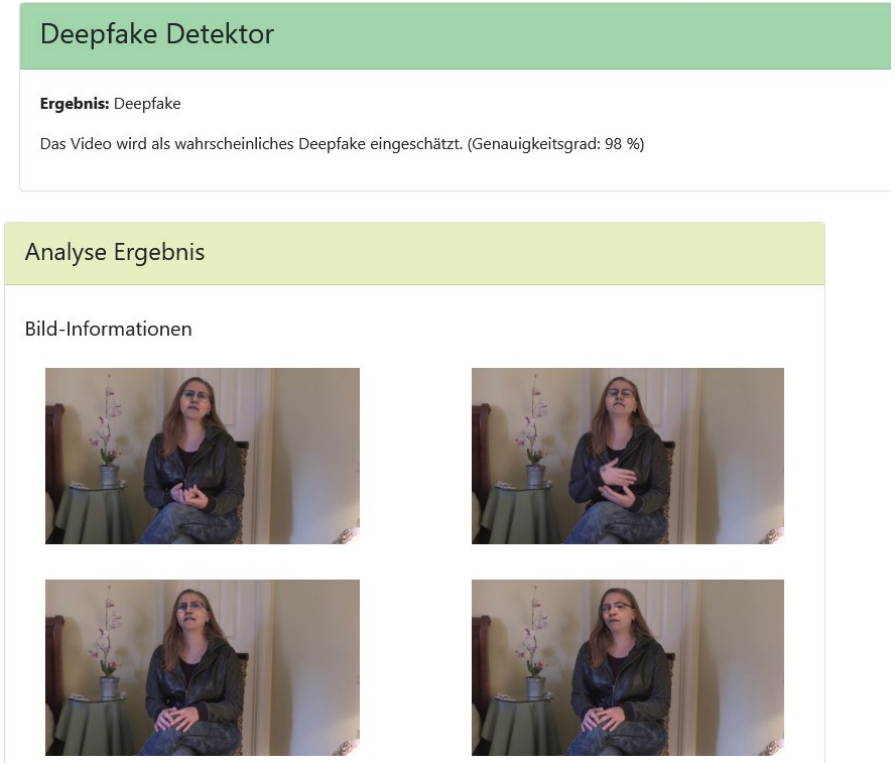
Image-based Geolocation Estimation

- **Task:**
  - Estimation of the recording location based on the available image information
- **Threat Scenarios:**
  - Photo manipulated to claim to have been taken in another country
- **Method:**
  - Estimate geo-location from image content
  - Deep Learning based pattern rechognition
  - 100km accuracy is often sufficient to verify the country or region.



Deep Fake Detection

- **Task:**
  - Detect manipulated Videos
  - Detect generated content within videos
- **Threat Scenarios:**
  - All disinformation types and categories
- **Method:**
  - Deep learning model to detect if a video or parts of it have been manipulated/generated
  - Visualization of modified/generated image parts



Detecting Extremist Symbols

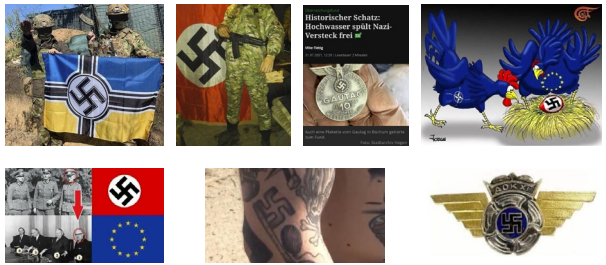
- **Task:**
  - Detect symbols of extremist groups
- **Threat Scenarios:**
  - Re-activation, prohibition law, radicalization, propaganda
- **Method:**
  - Deep Learning based object detection
  - Automated detection

Sun wheel

Wolfangel

Celtic cross

Swastika



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

TEXT BASED ANALYSIS

Natural Language Processing

FAKE NEWS DETECTION: TEXT CONTENT

- **Classification task**
  - Binary or multi-classification
- **Regression task**
  - Output as a numeric score of truthfulness

- **Similar NLP concepts**
  - Fake product reviews
  - Online resumes
  - Opinion spamming
  - Fake profiles
  - Spamming and phishing

CONTENT-BASED APPROACHES

- Headline & body text of articles
- Short comments (social media)

- **Linguistic cues and patterns**
  - Character, word, sentence or document level
  - Linguistic Inquiry and Word Count (LIWC)

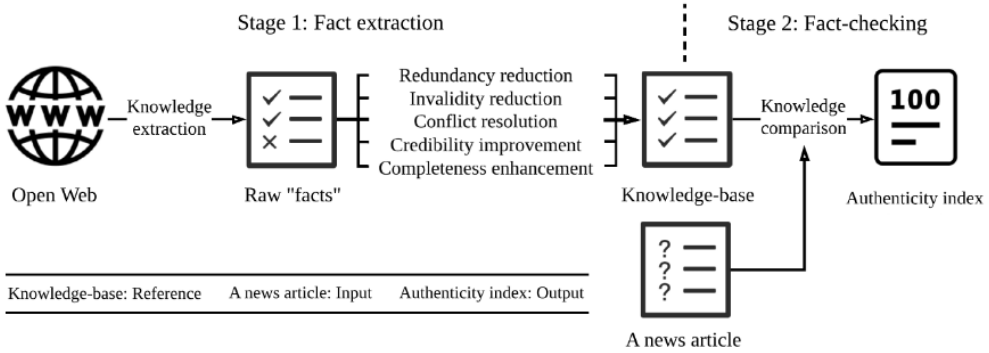
- **Style-based**
  - Hashtags, mentions, punctuation marks, sentiment
  - Topics, languages, domains

Attribute Type	Feature
Quantity	Character count
	Word count
	Noun count
	Verb count
	Number of noun phrases
	Sentence count
	Paragraph count
	Number of modifiers (e.g., adjectives and adverbs)
Complexity	Average number of clauses per sentence
	Average number of words per sentence
	Average number of characters per word
	Average number of punctuations per sentence
Uncertainty	Percentage of modal verbs
	Percentage of centainty terms
	Percentage of generalizing terms
	Percentage of tentative terms
	Percentage of numbers and quantifiers
	Number of question marks
Subjectivity	Percentage of subjective verbs
	Percentage of report verbs
	Percentage of factive verbs
	Percentage of imperative commands

Knowledge-based approaches

- Check for truthfulness of claims

- **Manual fact-checking**
  - Expert-based
  - Crowdsourced
  - politifact.com, snopes.com



- **Automatic fact-checking**
  - Linked Open Data (i.e. DBpedia)
  - Fact-extraction (Knowledge base construction)
  - Fact-checking (Knowledge comparison

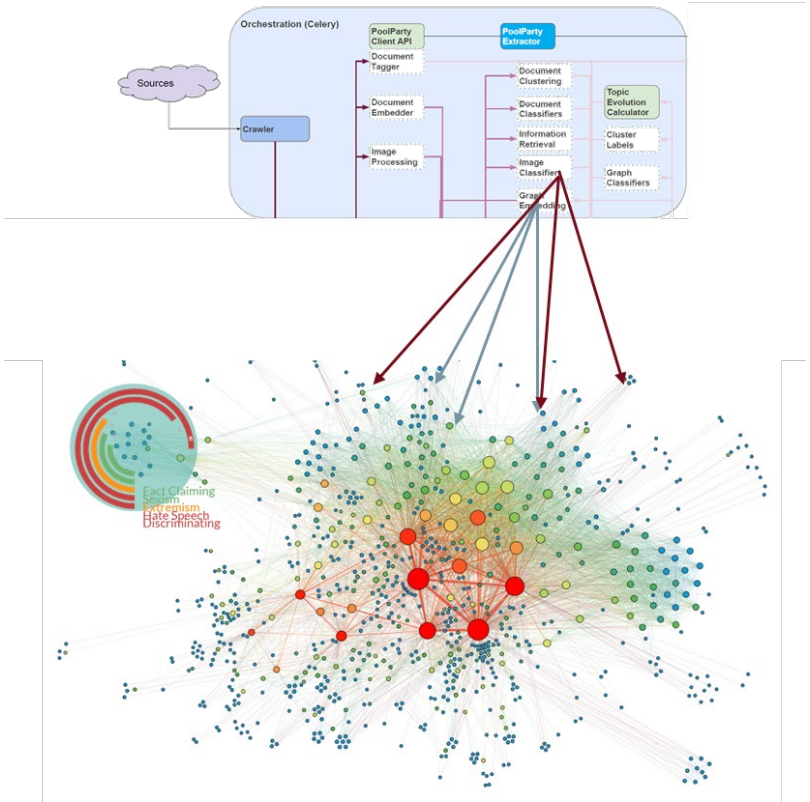
SOCIAL-CONTEXT Based APPROACHES

- **Stance-based**
  - Stance of body text relative to the headline claim
  - Viewpoint of user
  - Infer validity of original article
  - Support or refute claim
- **User-based**
  - Registration age
  - Numer of followers / followees
- **Propagation-based**
  - Propagation networks
  - i.e. Twitter shares / retweets
  - Likes
- **Credibility-based**
  - Headlines (clickbaits)
  - News source, spreaders, author

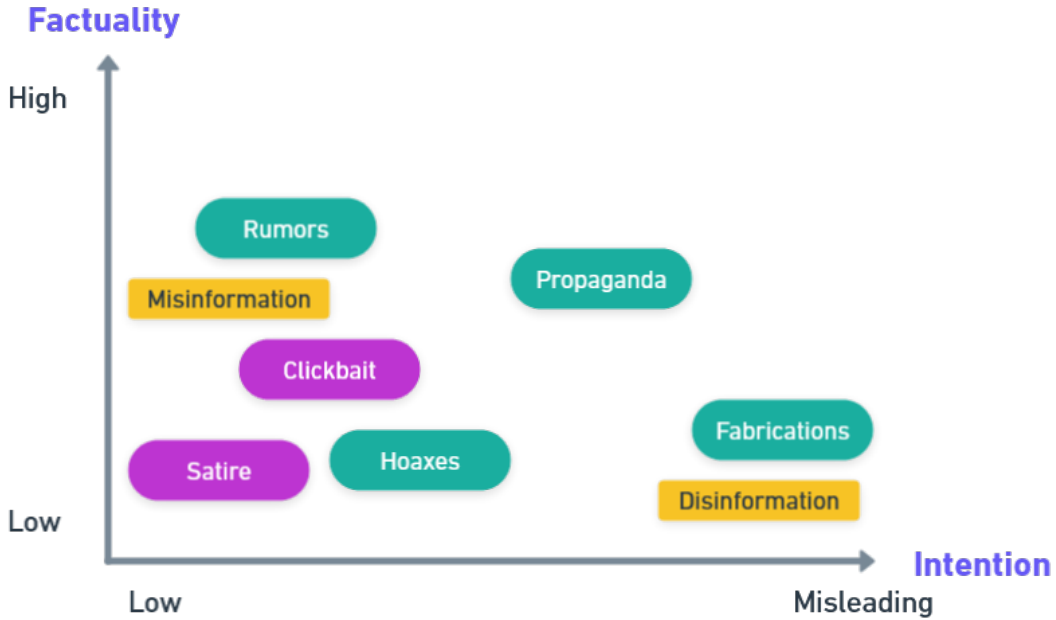
Indirect Detection of Disinformation

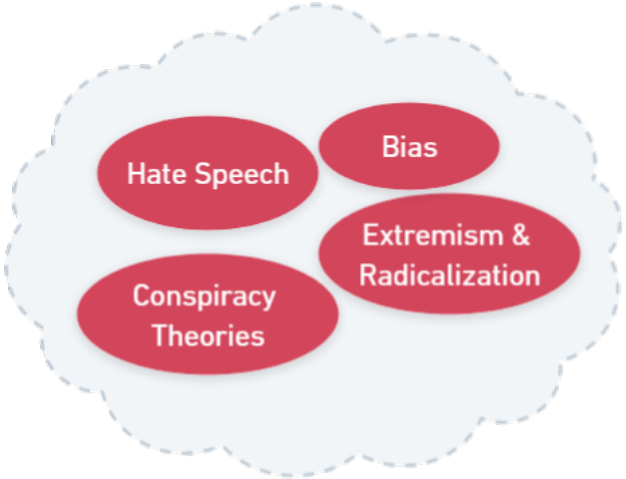
for Disinformation Analysis (in general)

- **Meta-Data Enrichment**
  - Specific machine learning models
  - detect different aspects / semantics of written content
    - E.g., writing style, reporting style, claims, narratives
- **Knowledge representation / modelling**
  - Linking of inferred information / meta-data
  - Semantic modelling
- **High-Level Inference**
  - Semantic approaches
    - E.g., ontologies
  - Graph based approaches
    - Graph analytics
    - Graph Neural Networks (GNN)



Fake news are intentionally and verifiably false?





Coverage of a wide range of semantic concepts

AUSTRIAN INSTITUTE OF TECHNOLOGY

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
Claim Detection	Were claims stated?	German	Social networks	Yes / No
Appealing contents	Appealing, positive, discussion-promoting, language	German	Social networks Article	Yes / No
Sentiment	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 criminally relevant?	German	Social networks	Incitement, insult
Sexism	Various categories of sexism	English	Social networks	Misogyny, Sexual Violence

DeTox

Toxicity & Hate Speech Detection

Collection of **German Tweets** via TwitterAPI

781,991 Tweets from
154,151 Twitter users

First half of 2021

Controversial Topics that were trending at that time

- „Querdenker“
- „Corona“
- …

Special Focus on Conversations for Toxicity Analysis

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics

DeTox

Toxicity & Hate Speech Detection

- 10,278 Labeled Comments
- 6 Annotators
  - Annotation Guidelines
  - Bi-Weekly Discussion Meetings
  - Inter Annotator Agreement
- **10 machine learning models**

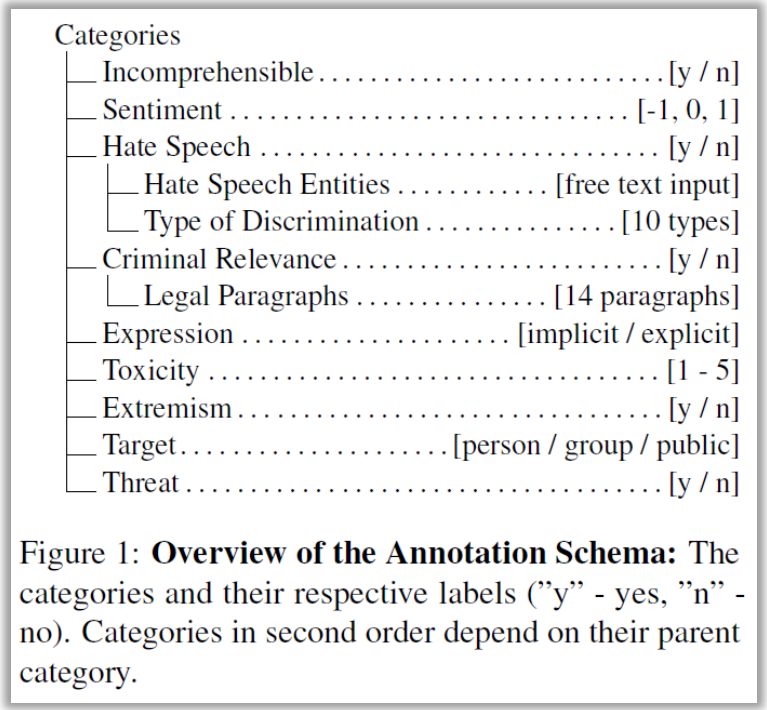
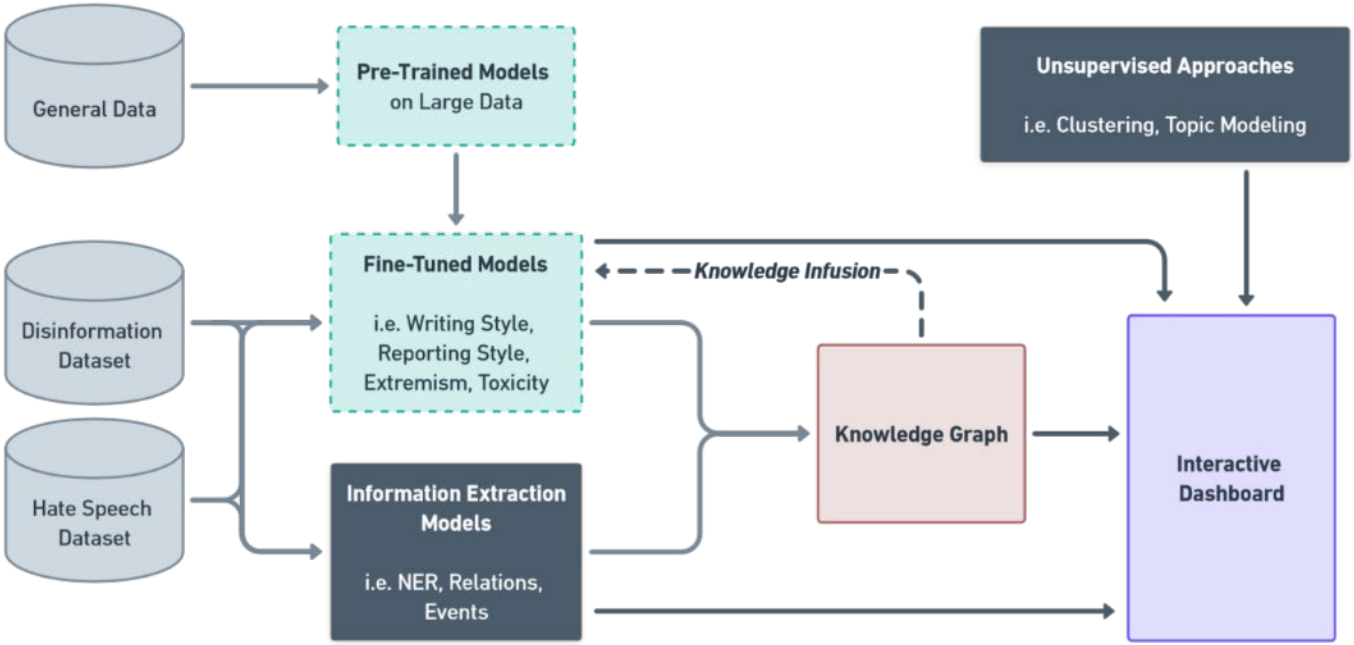


Figure 1: **Overview of the Annotation Schema:** The categories and their respective labels (”y” - yes, ”n” - no). Categories in second order depend on their parent category.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics

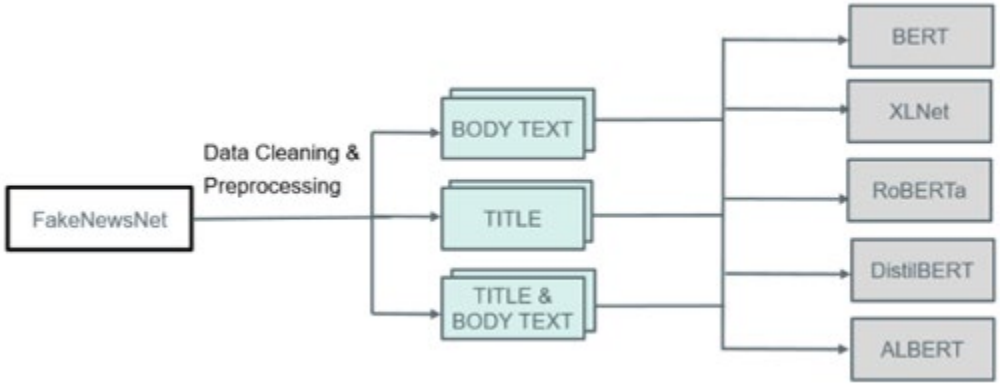
Classification Approach

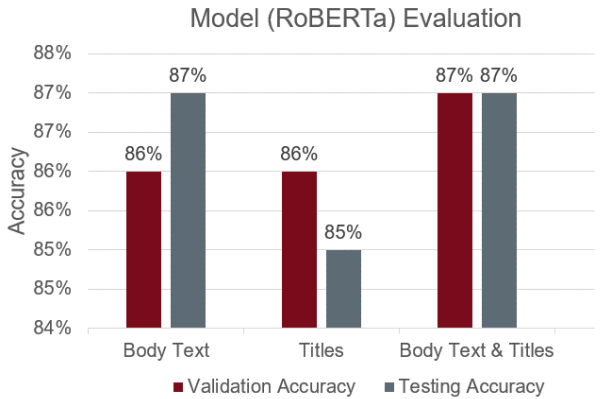


Sch¨utz, M.: Disinformation detection: Knowledge infusion with transfer learning and visualizations. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. pp. 468–475. Springer Nature Switzerland, Cham (2023)

Fake News –Module 1

- **Type: binary classification**
  - *Fake News, Not Fake*
- **Language:** EN
- **Data:** news articles
  - (FakeNewsNet Dataset)
- **Count:** ca. 22.000





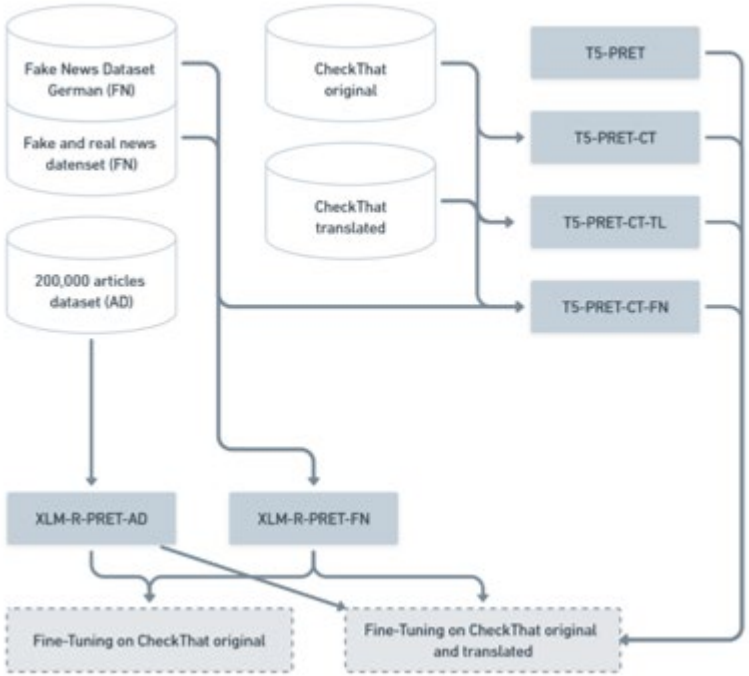
M. Schütz (2021). "Detection and Identification of Fake News: Binary Content Classification with Pre-trained Language Models"; in: "Information between Data and Knowledge - Schriften zur Informationswissenschaft", 74; W. H ülsbuch (ed.); vwh Verlag, Glückstadt, Deutschland, 2021, ISBN: 978-3-86488-172-5, 422 - 43

M. Schütz, A. Schindler, M. Siegel, K. Nazemi: "Automatic Fake News Detection with Pre-Trained Transformer Models"; in: "Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science", 12667; Springer, Cham, 2021, ISBN: 978-3-030-68786-1, 627 - 64

Fake News – Module 2

Multiclass

- **4 classes:**
  - *Fake News, partly Fake News, No Fake News, Misc*
- **Language: GE & EN**
- **Data: News-Articles**
  - (CheckThat Shared Task 2022)
- **Number:** ca. 2.000
- **Pre-trained with 200.000 articles**

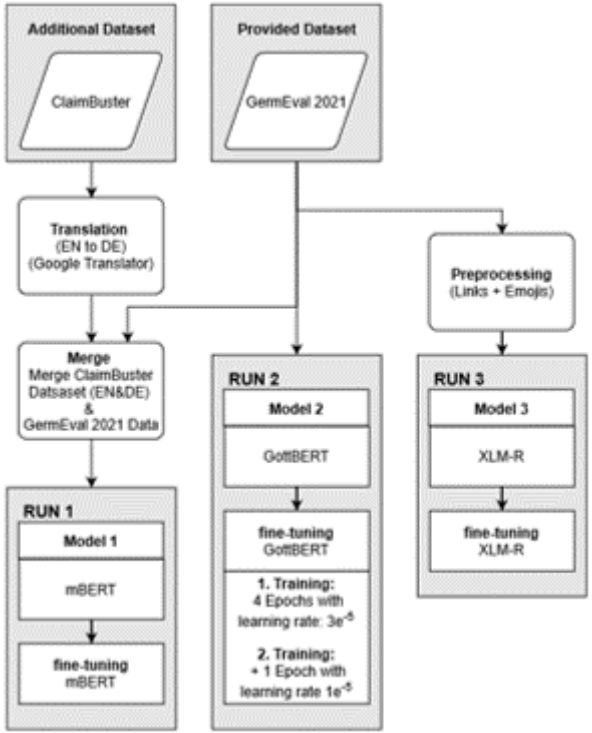


- **Note: also provided baseline for the shared task!**

M. Schütz, J. Böck, M. Andresel, A. Kirchknopf, D. Liakhovets, D. Slijepcevic, A. Schindler (2022). AIT FHSTP@CLEF2022 Check-That! Lab: Cross-Lingual Fake News Detection with a Large Pre-TrainedTransformer.

Claim Detection

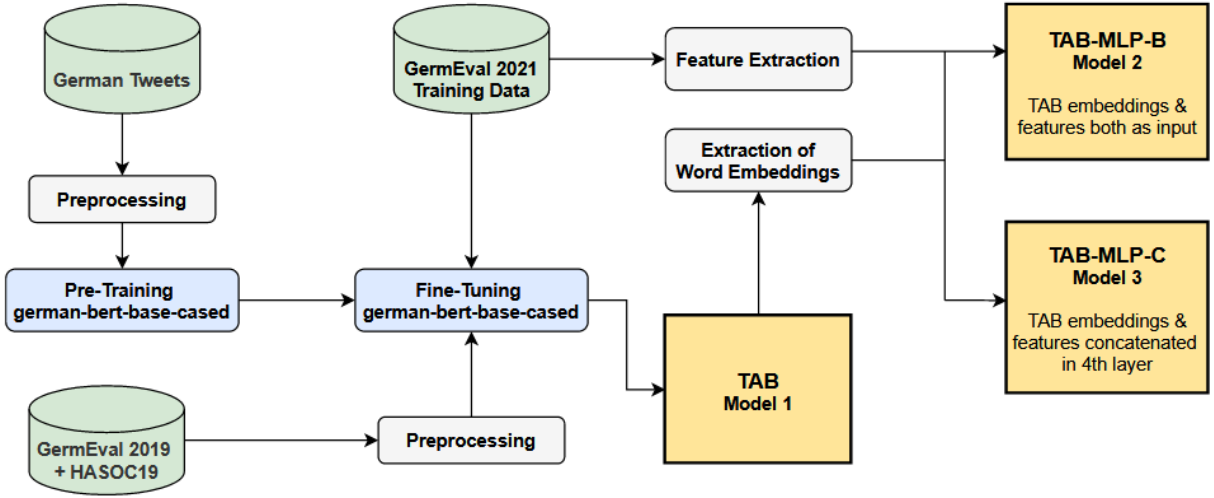
- **Type: binary classification**
  - *Claims, No Claims*
- **Language:** German
- **Data:** Social Media (GermEval 2021)
- **Count:** ca. 3.300
- **Multi-lingual transfer learning**



J. Boeck, D. Liakhovets, M. Schütz, A. Kirchknopf, D. Slijepcevic, M. Zeppelzauer, A. Schindler (2021). AIT FHSTP at GermEval 2021: Automatic Fact Claiming Detection with Multilingual Transformer Models. In: Proceedings of the GermEval 2021 Workshop on the Identification of Toxic, Engaging, and Fact-Claiming Comments, netlibrary, Düsseldorf, Germany, 6 pages

Toxic comment Detection

- **Binary classification:** *Toxic, Not Toxic* **Language:** German
- **Data:** Social Media (GermEval 2021) **Count:** ca. 3.300



Feature	Toxic	Not Toxic
word count	201	179
punctuation count	7.41	6.84
exclamation count	0.69	0.31
question mark count	0.48	0.36
word punctuation ratio	0.0111	0.0138
word exclamation ratio	0.0027	0.0021
word question mark ratio	0.0020	0.0030
hate word count	0.32	0.24
hate word count ratio	0.0017	0.0014
character capslock ratio	0.0306	0.0168
sentiment	-0.0147	-0.0080
emoji count	0.49	0.13
emoji sentiment	0.0424	0.0191
word emoji ratio	0.0457	0.0227

Mina Schütz, Christoph Demus, Jonas Pitz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2021. DeTox at GermEval 2021: Toxic Comment Classification. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 54–61, Duesseldorf, Germany. Association for Computational Linguistics.

Information Extraction Tasks for Disinformation

- **Extraction of claims and statements in text**
- **Extraction of entities and their relationships in text**
- **Event detection**
- **Keyword extraction**
- **Topic Modeling**
- **Clustering Approaches**

Information Extraction

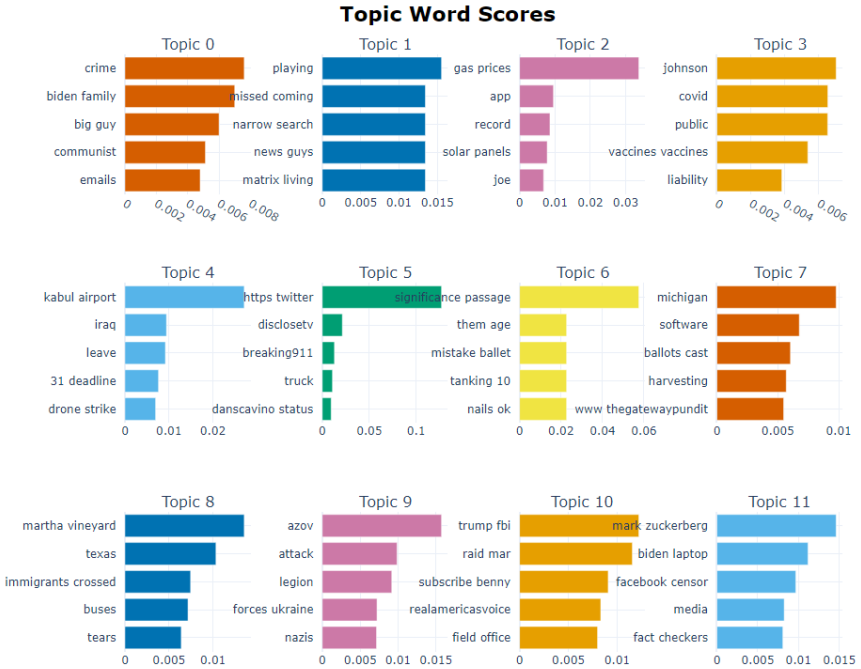
Named Entity Recognition & Topic Modeling

AUSTRIAN INSTITUTE OF TECHNOLOGY

- **Task:**
  - quick overview of the text (places, organizations, people, other)
  - Link articles, messages, etc. by events, locations, etc.

Ein Gastbeitrag von Prof. Dr. Thomas Rießinger PER „ Der Versuch, den Himmel auf Erden MISC zu errichten“, schrieb der Philosoph Karl Raimund Popper, PER „erzeugt stets die Hölle“, genauer gesagt: „eine jener Höllen, die Menschen für ihre Mitmenschen bereiten“. In unseren glücklichen Tagen sollte diese Äußerung die Aufmerksamkeit des so trefflich von Thomas Haldenwang PER geleiteten Bundesamtes für Verfassungsschutz ORG erregen, denn an staatlichen Versuchen, uns den Himmel auf Erden zu bereiten, fehlt es nicht und Popper PER macht sich hier eindeutig der verfassungsschutzrelevanten Delegitimierung des Staates schuldig. Am klimatischen Himmel wird intensiv gearbeitet, das himmlische Zeitalter der Elektromobilität soll uns vor dem Untergang retten, die kontaminierten und toxischen Sprachgewohnheiten alter weißer Männer sollen einem himmlisch-gerechten Sprachregime weichen – aber vor allem führt uns die Elite zu einem wahren und bisher unerreichten Himmel der Gesundheit. Krankheiten oder gar Viren MISC darf es in einer himmlisch organisierten Gesellschaft nicht mehr geben und das Mittel der Wahl ist

- **Task:**
  - Finding trending topics over time, which contain disinformation

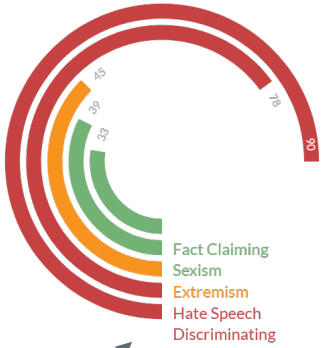


Information Nutrition Labels

Similar to food nutrition labels

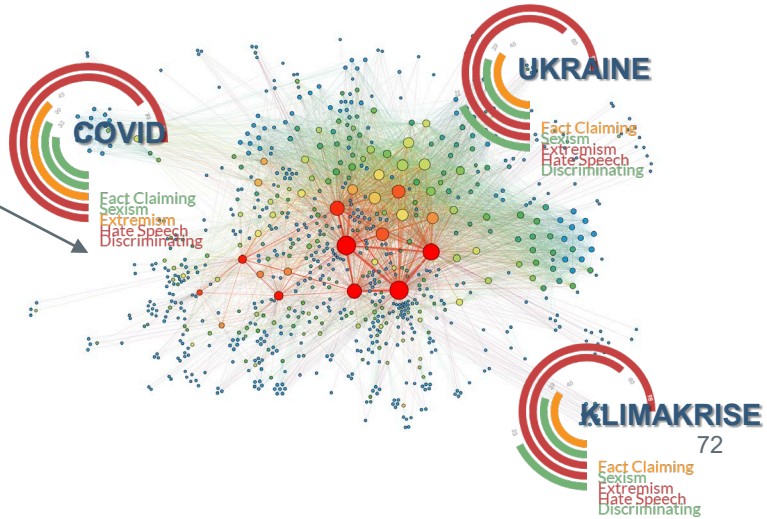
Describe the content of documents or online articles
in a clear understandable form

Users get a quick assessment of the information content.



AI models for content description

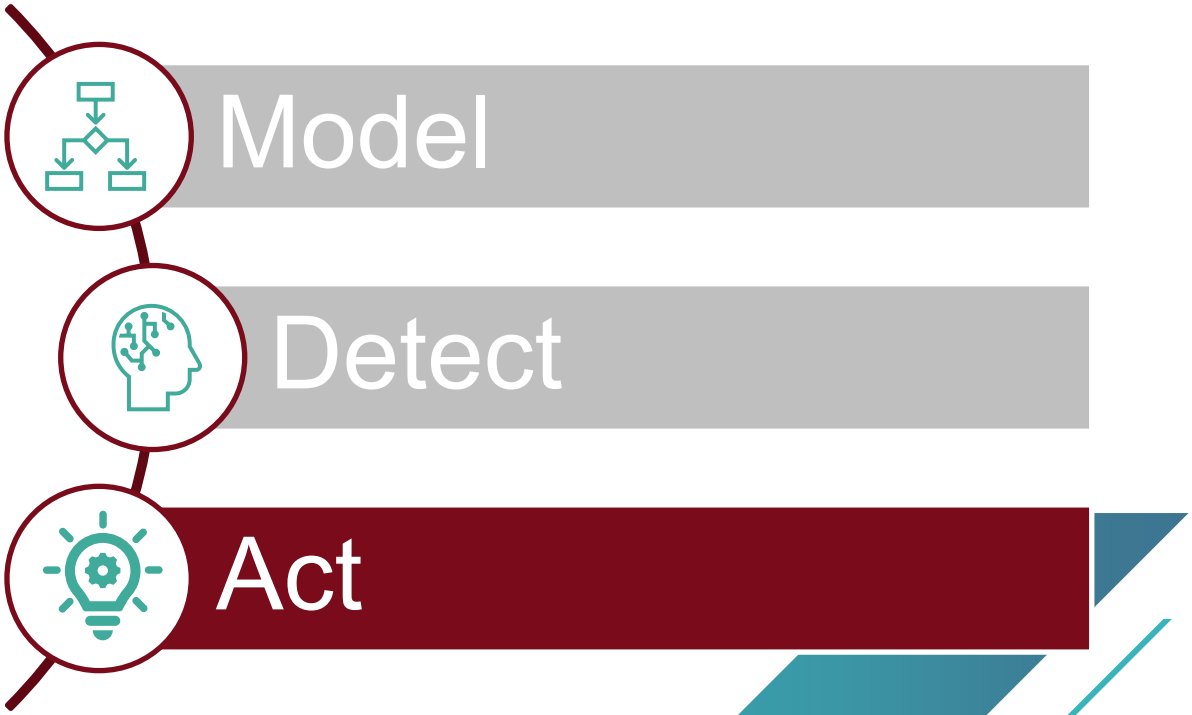
Name	Erkannte Inhalte	Sprache	Domäne	Kategorie Beispiele
Fake News	Direkt Erkennung von Fake News	Englisch	Soziale Netzwerke	Ja / Nein
Hassrede	Hass gegen Gruppen oder Individuen	Multi-ling	Soziale Netzwerke Diskussionsforen	Ja / Nein
Extremismus	Extremisitische Inhalte	Deutsch	Soziale Netzwerke Artikel	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



Countering Fake News & Disinformation

How can we… ?

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

SITUATIONAL AWARENESS

Detecting Fake News Trends

Extremisten instrumentalisieren COVID-Proteste (Motivation RAIDAR)

Auch in Social Media

„Kritiker.at“ (Fiktiver Kanal)
- *selbst-organisierter Kanal*
- *Ausdruck und zur Kommunikation der Unzufriedenheit*
- *Protests gegenüber dem Status-Quo*
- *„Wutbürger“*

„Radikale.at“ (Fiktiver Kanal)
- *organisierte radikal-extremistische Vereinigung*
- *propagandistischen Ziele*

https://www.derstandard.at/story/2000133302828/corona-demos-unter-gefaehrlichen-flaggen

Forschungsfragen

Grundlegende Einschätzung der Rhetorik, Grad der Radikalisierung

(F1) Ist Kritiker.at schon radikalisiert, und, wenn ja, zu welchen Grad?

(F2) Welche Taxonomie, Argumente, Narrative werden in Kritiker.at verwendet?

(F3) Können Aussagen, bzw. Formulierungen in Kritiker.at schon Paragraphen im Kontext von Hass im Netz zugeordnet werden?

In Bezug auf Interventionsmaßnahmen

(F4) Kann eine Reaktion auf Interventionsmaßnahmen festgestellt, gemessen werden?

(F5) Kann die Reaktion auf Interventionsmaßnahmen interpretiert, quantifiziert werden?

(F6) Sind Taxonomien, Argumente, Narrative in Kritiker.at gleich verteilt, oder gibt es Gruppierungen, bzw. wie sind diese verteilt?

(F7) Welchen Einfluss haben diese Gruppen auf einander?

(F8) Wie verändern sich Gruppenverteilung und Einfluss über die Zeit? Können hier Tendenzen einer Radikalisierung oder Deradikalisierung festgestellt werden?

Einfluss von bekannten radikalen oder extremistischen Gruppierungen zu erkennen / messen

(F9) Welche Taxonomie, Argumente, Narrative werden in Radikale.at verwendet?

(F10) Welche Überlappungen bei Taxonomien, Argumenten, Narrativen können zwischen den beiden Domänen festgestellt werden?

(F11) Kann eine Beeinflussung, bzw. Radikalisierung von Kritiker.at durch Radikale.at festgestellt werden?

Verstöße hinsichtlich für Hass im Netz relevanter Paragraphen

(F12) Können Inhalte der von Kritiker.at und Radikale.at automatisiert Paragraphen zugewiesen werden (ohne automatisierte Entscheidung über Verstoße)?

(F13) Kann die automatisierte Identifikation relevanter Text-Passagen nachvollziehbar erklärt und dargestellt werden?

Big Data Problem

- Example: **Impfschaden_D_AUT_CH**

- Downloaded messages : 35.366
- Messages with Text: 22.854
- Number of words: 1.210.170

- Downloaded articles: 753
- Number of words: 606.103

- Total number of words : 1.824.300

- ~ 18 Fantasy Books
- Time required to read: 1.824.300 / 200 / 60 / 38.5 ➔ 4 work weeks
  - (word count / Words per minute / min per hour / work-hours per week)

- Downloaded images: 3.039

Word counts of the books in J.R.R. Tolkien's *Lord of the Rings* series:

- *The Hobbit* – 95,356 words
- *The Fellowship of the Ring* – 187,790 words
- *The Two Towers* – 156,198 words
- *The Return of the King* – 137,115 words
- The entire *Lord of the Rings* series (including *The Hobbit*) – 576,459 words

Word counts of the books in J.K. Rowling's *Harry Potter* series:

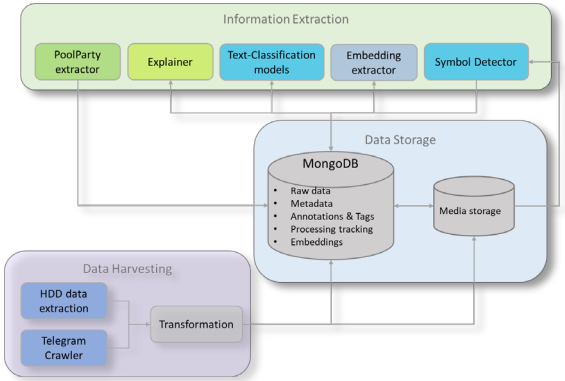
- *Harry Potter and the Philosopher's Stone – 76,944 words*
- *Harry Potter and the Chamber of Secrets – 85,141 words*
- *Harry Potter and the Prisoner of Azkaban – 107,253 words*
- *Harry Potter and the Goblet of Fire – 190,637 words*
- *Harry Potter and the Order of the Phoenix – 257,045 words*
- *Harry Potter and the Half-Blood Prince – 168,923 words*
- *Harry Potter and the Deathly Hallows – 198,227 words*
- *The entire Harry Potter series – 1,084,170 words*

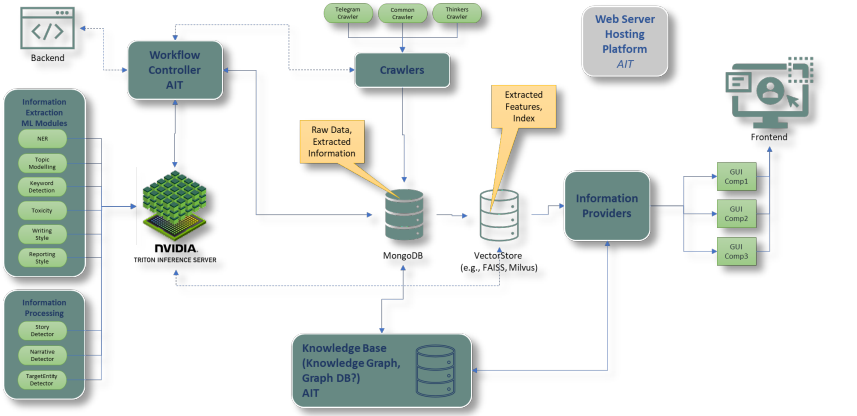
Word counts of C.S. Lewis's The Chronicles of Narnia series:

- *The Lion, The Witch, and the Wardrobe – 38,421 words*
- *Prince Caspian – 46,290 words*
- *The Voyage of the Dawn Treader – 53,960 words*
- *The Silver Chair – 51,022 words*
- *The Horse and His Boy – 48,029 words*
- *The Magician's Nephew – 64,480 words*
- *The Last Battle – 43,333 words*
- *The entire Chronicla series – 345,535 words*

Large Scale Analysis Platforms

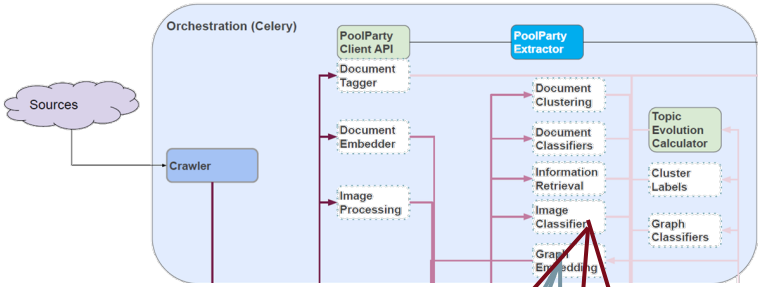
- **Input channels**
  - Social Media / general content Crawlers
  - Hard disk data (case related)
- **Processing chains (pipelines)**
  - Text classification
  - Text similarity determination
  - Semantic text analysis
  - Image recognition
  - Explainable AI
- **Efficient data processing**
  - Large Scale Processing
  - GPU Computing
  - Resource Management
  - High Performance Computing



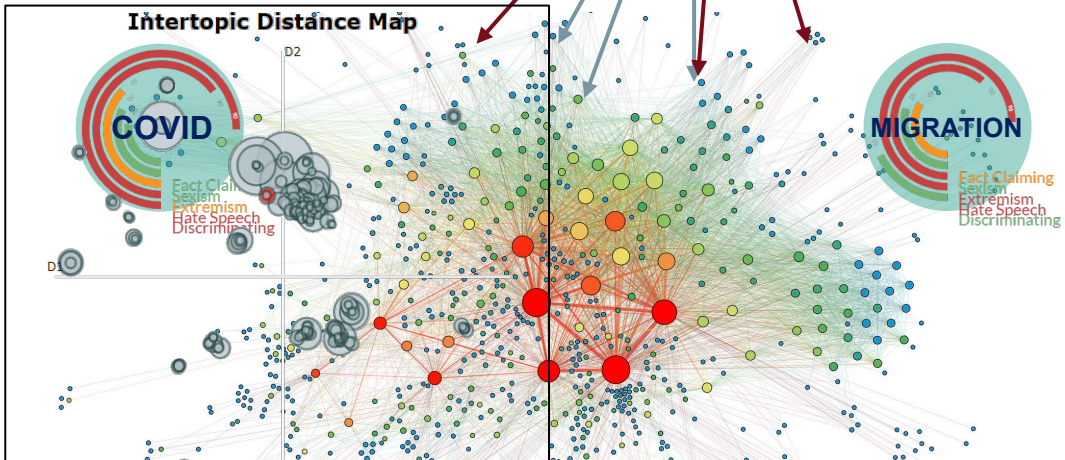


Multi-Modal Information Fusion

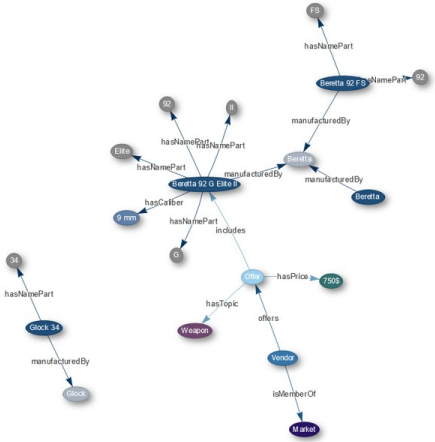
Integrated Micro-Services Architecture



Cluster / Topic Modelling



Relationship Extraction



Goals

- Effective integration of extracted information to address use-case requirements
- Answer research questions

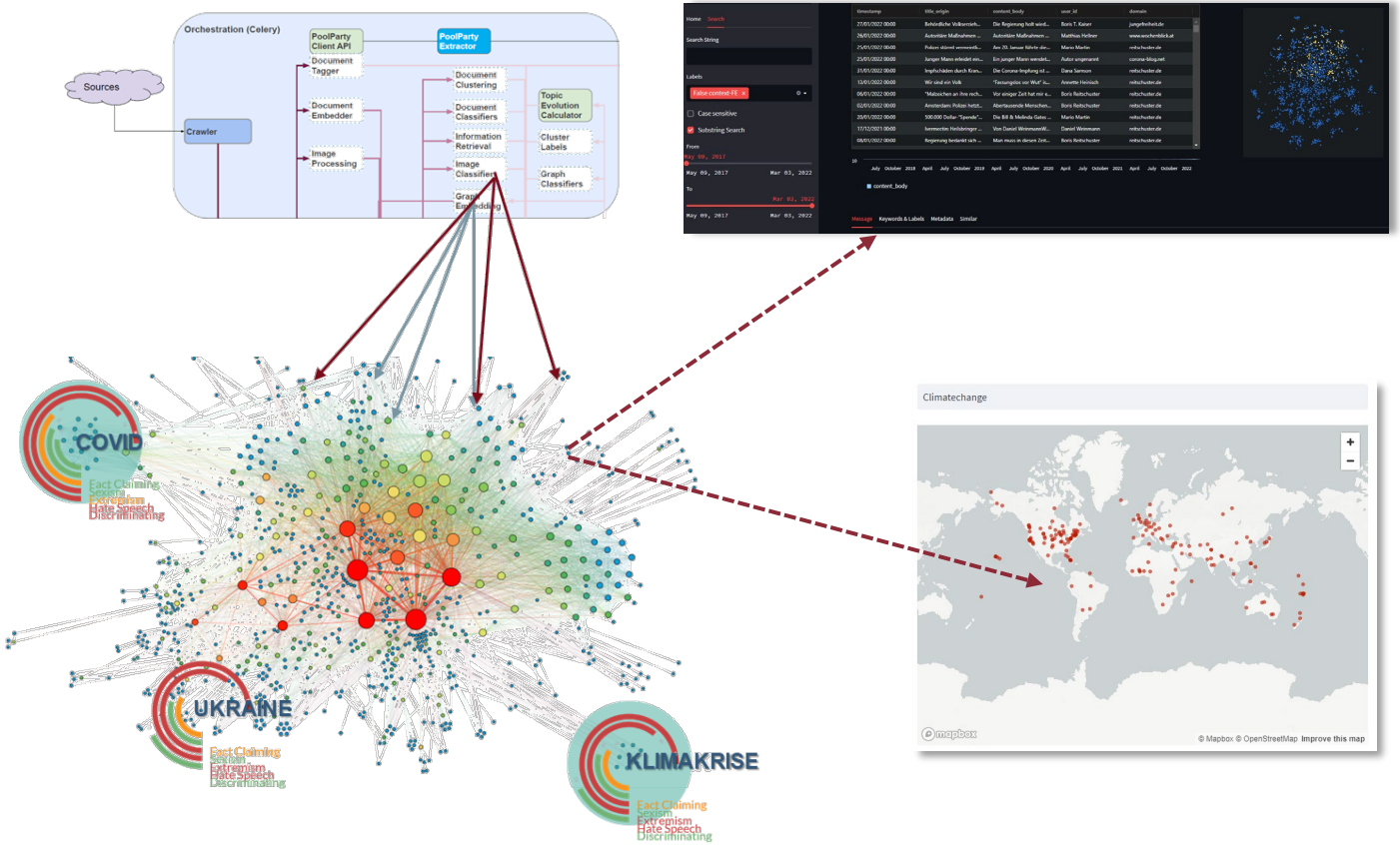
Use-Cases

- Desinformation Analysis
- Radical / Extremist Content Detection
- Hate Crime Detection
- Large Scale content analysis
- OSINT

Methods

- Graph modelling / analytics
- Link predictions
- Relationship extraction
- Graph embeddings
- Knowledge modelling
- Hybrid Systems

AI-Based trend analysis



Goals

- Present / visualize fused information to create knowledge
- Further Information aggregation to provide high-level answers to societal challenges

Use-Cases

- Novel approaches to search and discovery
- Knowledge mapping
- Complex automated trend analysis (at early stages)

Projects

- Fusion of embedding spaces
- Maximum Marginal Relevance Retrieval
- Search space visualization

Rapid Content Summarization

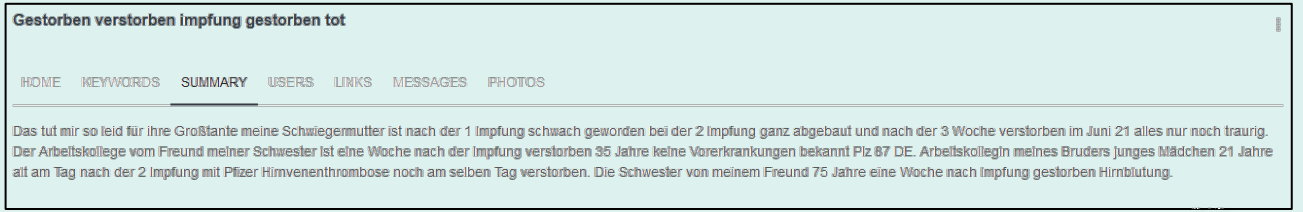
Questions:

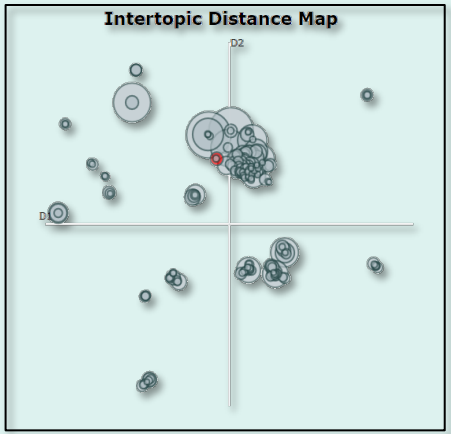
- What discussions are taking place in a particular social media channel?

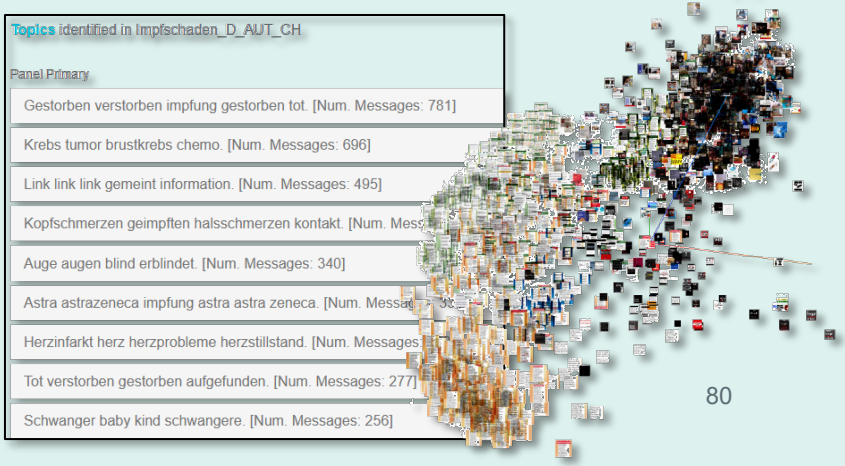
Methods:

- Topic Modelling
- Automatic Key-term Detection
- Automatic Summarization
- Near Duplicate Detection / Story Clustering
- Named Entity Recognition

- Narrative-Storyline Analysis
  - ➔ Trend Analysis
  - Origin of Campaign
  - Spatio-Temporal Distribution
  - “Weather Radar for Fake News”



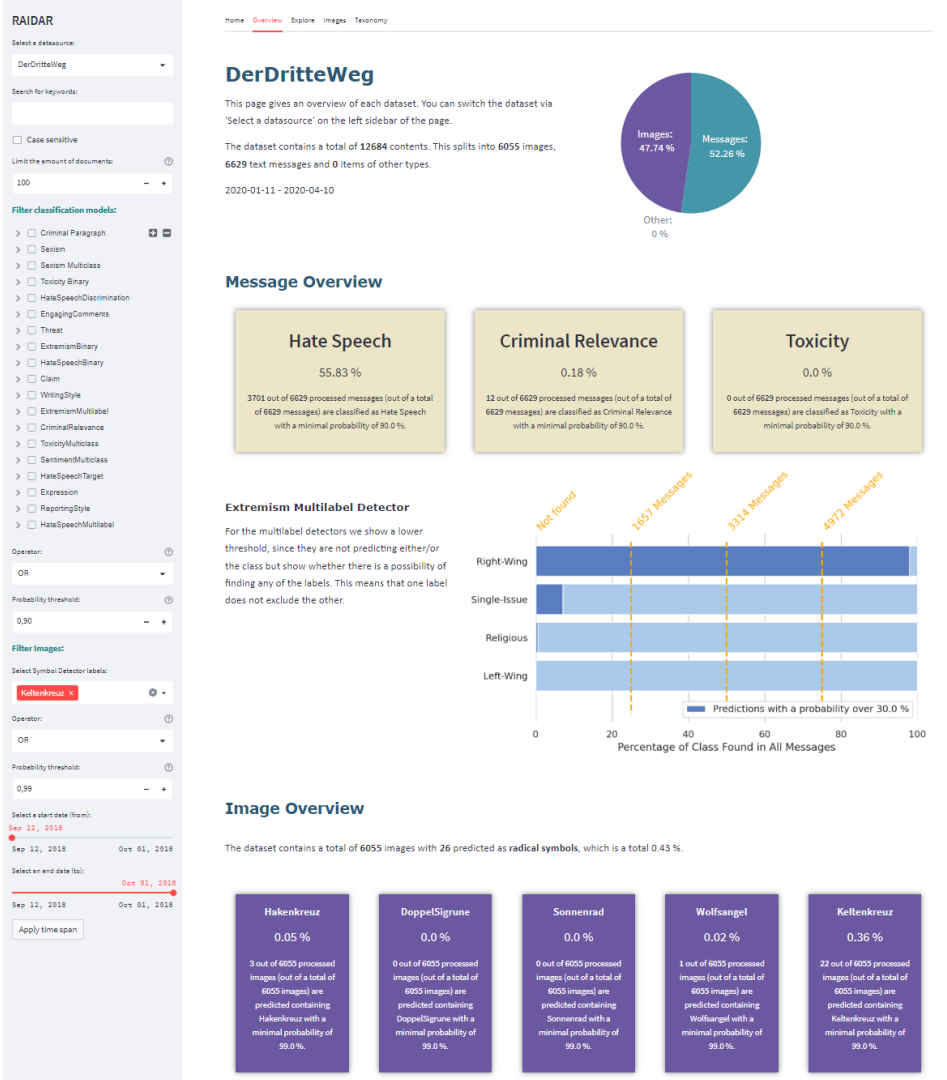




Content Assessment / Evaluation

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



Exploratory search

AUSTRIAN INSTITUTE OF TECHNOLOGY

RAIDAR

Select a datasource:

expedition_avantura

Search for keywords:

Case sensitive

Limit the amount of documents:

20

Filter classification models:

- Criminal Paragraph
- Sexism
- Sexism Multiclass
- Toxicity Binary
- HateSpeechDiscrimination
- EngagingComments
- Threat
- ExtremismBinary
- HateSpeechBinary
- Claim
- WritingStyle
- ExtremismMultilabel
  - Left-Wing
  - Religious
  - Right-Wing
  - Single-Issue
- CriminalRelevance
- ToxicityMulticlass
- SentimentMulticlass
- HateSpeechTarget
- Expression
- ReportingStyle
- HateSpeechMultilabel

Operator:

OR

Probability threshold:

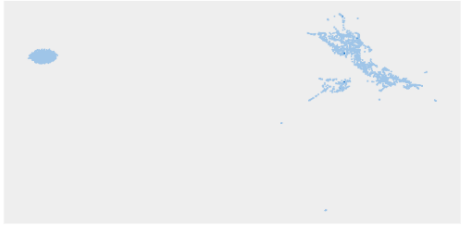
0,80

Filter Images:

Home Overview Explore Images Taxonomy

Explore the data

Here you can explore the datasource you selected. Each point in the scatterplot represents a document in the dataset. The closer the documents are to each other, the more similar the contents of the documents. With the so-called 'lasso' function you can select a group of documents. The highlightes documents are the filtered ones.



5 Search Results

The table shows all documents that were filtered and selected. If you click on one message, you can read the message below in the 'Message' tab.

id	date	text
645e0bd008ccbdf7d9330eda	27/02/2022 15:38	@boris_rozhin @grey_zone @go338 @omonmos
645e0ea108ccbdf7d9333a7f	31/03/2022 15:55	#rechtspopulisten
645e0f2b08ccbdf7d9334296	08/04/2022 04:53	Wenn die Pervers-Gruppe-Deluxe Ramstein ein Vi
645e0f7008ccbdf7d93346c7	14/04/2022 14:36	Putin "Fuck your denazification!" Eu We sup
645e0f8a08ccbdf7d9334861	17/04/2022 07:40	Eine Duginistin und "Nationalbolschewistin".

Documents over time

Additional information:

Message Labels Metadata Attachments Similar

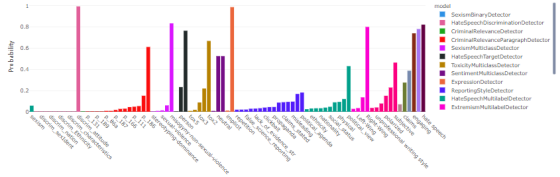
Wenn die Pervers-Gruppe-Deluxe Ramstein ein Video übers Älterwerden und "Schönheits"-OPs macht und versucht mit einem kleinen Seitenhieb auf Transgender-Geschlechtsumwandlungen wieder für neue PR-Gags zu sorgen: Lol. Es gibt im übrigen kaum ein Land wo diese antideutsche und dekadente Zeckengruppe so beliebt ist wie Russlan.
https://youtu.be/hBTNuJ3ILWI

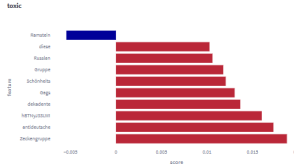
Classifications

In this section the labels are shown that the models predicted. You can adjust the threshold to only show predictions with a certain probability. The threshold is always a value between 0 (not predicted label) and 1 (predicted label). The box shows the name of the detection model, the found label and its prediction probability.

Select a probability threshold:

0,80





Wenn die Pervers - Gruppe - Deluxe Ramstein ein Video übers Älterwerden und " Schönheits " - OPs macht und versucht mit einem kleinen Seitenhieb auf Transgender - Geschlechtsumwandlungen wieder für neue PR - Gags zu sorgen : Lol . Es gibt im übrigen kaum ein Land wo diese antideutsche und dekadente Zeckengruppe so beliebt ist wie Russlan . https :// youtu . be / hBTNuJ3ILWI

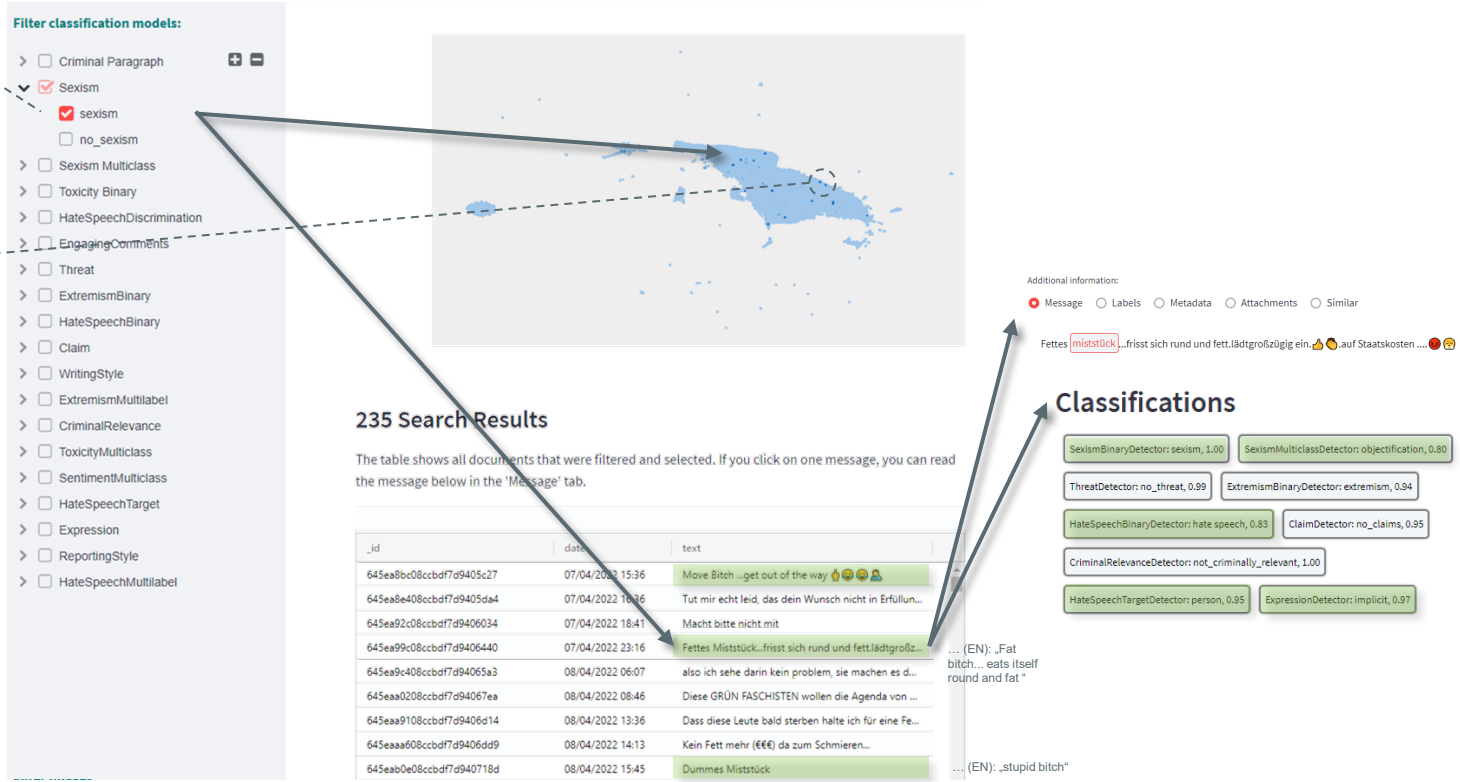
Filter by AI Module Results

Filter by AI Module result

- Select on or many filters
  - Combinations
    - OR – inclusive / union
    - AND – excluse / intersection

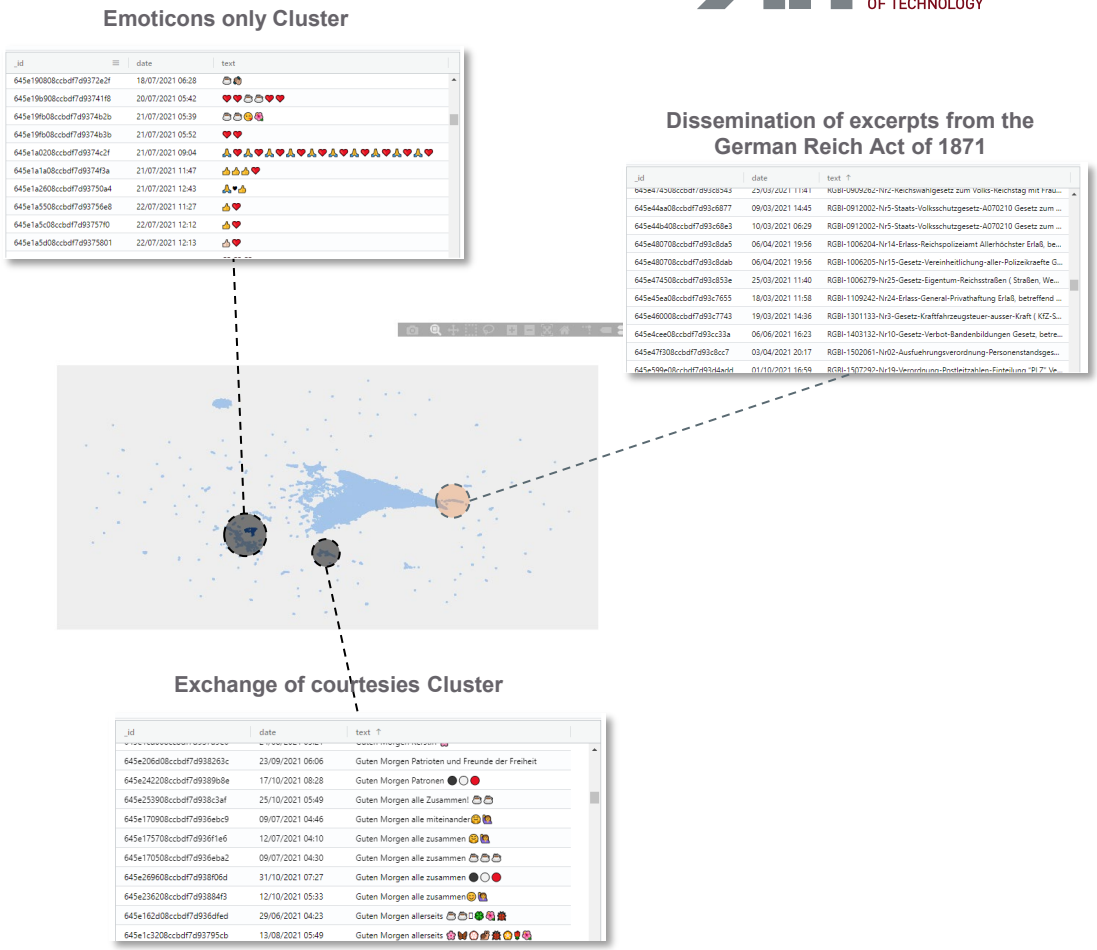
Visualize Results

- Identify accumulations



Visual Exploration

- **Task:**
  - Group similar content together
  - Visualize content similarities
- **Objective:**
  - Assistance in identifying/distinguishing between relevant and irrelevant content
  - Content overview/understanding
- **Method:**
  - Content Embeddings
    - E.g., text-, image-, audio-embeddings
  - Graph Embeddings
  - Projections
    - E.g., t-sne, umap



Analytical Results

Document Collection: Telegram Channel „Eiserne_Jugend“

Number of Documents in Cluster/Topic: 15

Generated Title:

Der Krieg und die Überwindung der Ideenwelt der Französischen Revolution

Keywords:

Extreme jüdische Intellektualismus, **Deutsche Revolution**, **Deutscher Weg**, Bücher, **Krieg**, **Wehrmacht**, Oberbefehl, **Adolf Hitler**, Reform, Französische Revolution, **Rasse**, Feigling, Unglückszeiten, Stern, Dunkle Wolken, Welt, Anblick, Unterwerfung, Verdruss, Widerstand

Extracted Key-Claims / Statements:

1. Der **extreme jüdische Intellektualismus ist nun zu Ende**.
2. Der **deutsche Weg ist frei**, da der Durchbruch der deutschen Revolution stattgefunden hat.
3. Der zukünftige deutsche Mann wird nicht nur ein Mann der Bücher sein, sondern auch ein Mann mit Erfahrungen und Fähigkeiten.
4. Der Krieg, der von der deutschen Wehrmacht unter dem Oberbefehl Adolf Hitlers geführt wird, ist ein Krieg von ungeheurer Reform.
5. Der **Krieg überwindet** nicht nur die **Ideenwelt der Französischen Revolution**, sondern **vernichtet auch** direkt all jene **rassischen** und menschenunwürdigen **Elemente**, die sich an der Wahl des jungen Adolf Hitlers **für ein Leben in München** widerspiegelten.

Narrative Analysis

- Die Narrative des **Nationalsozialismus**
  - *Diese Narrative beschreibt die Ideologie und Weltanschauung des Nationalsozialismus und die Ziele und Taten, die auf diese basieren.*
- Die Narrative von **Freiheit und Unabhängigkeit**
  - *Diese Narrative beschreibt die Bedeutung von Freiheit und Unabhängigkeit für das Volk und die Notwendigkeit, diese Werte zu verteidigen und zu fördern.*
- Die Narrative des **Vaterlands**
  - *Diese Narrative beschreibt die Bedeutung und Wert des Vaterlands für das Volk und die Notwendigkeit, es zu schätzen und zu schützen.*
- Die Narrative des **Blutes**
  - *Diese Narrative beschreibt die Bedeutung und Wert des Blutes des Volkes und die Notwendigkeit, es zu bewahren und zu schützen.*
- Die Narrative des **Judentums**
  - *Diese Narrative beschreibt die Rolle und Bedeutung des Judentums im nationalsozialistischen Deutschland und die Auswirkungen dieser Rolle auf das Volk.*

Diese Narrativen sind miteinander verflochten und haben eine große Bedeutung für die Einstellungen und Werte des Volkes. Sie bilden die Grundlage für die politische und gesellschaftliche Realität des Volkes und haben Auswirkungen auf die individuellen Lebensweisen und Überzeugungen der Menschen.

Further analysis

Analyse der Vorurteile

- **Antisemitismus**
  - Der Text enthält Vorurteile gegenüber Juden und deren Eigenschaften, wie zum Beispiel in der Aussage *"Der Jude ist nur einig, wenn eine gemeinsame Gefahr ihn dazu zwingt oder eine gemeinsame Beute lockt fallen beide Gründe weg so treten die Eigenschaften eines krassesten Egoismus in ihre Rechte und aus dem einigen Volk wird im Handumdrehen eine sich blutig bekämpfende Rotte von Ratten.*"
- **Rassismus**
  - Der Text enthält Vorurteile gegenüber anderen Menschen, die nicht der "Arier-Rasse" angehören, wie zum Beispiel in der Aussage "*Wären die Juden auf dieser Welt allein so würden sie ebensosehr in Schmutz und Unrat ersticken wie in haßerfülltem Kampfe sich gegenseitig zu übervorteilen und auszurotten versuchen*".
- **Nationale chauvinistische Ideologie**
  - Der Text enthält Vorurteile gegenüber anderen Nationen und Kulturen, wie zum Beispiel in der Aussage "*Der Restlose Mangel jedes Aufopferungssinnes kann den Kampf zum Theater werden lassen."*

Aufforderung zur Gewalt

- **In der Aussage** "*Man schmäht dies Wort darum das Schwert zur Hand Und dulde nicht dass man dies Wort entweiht*" wird eine **Aufforderung zur Gewalt gegen diejenigen** gegeben**, die das Wort Vaterland ablehnen oder entweihen**.
- Die Aussage "*Gott gab ein Stück von seiner EwigkeitUns Deutschen in dem Worte Vaterland*" legt nahe, dass die **Verteidigung des Vaterlands eine moralische Pflicht** ist, die **mit Gewalt durchgesetzt** werden muss.
- Die Aufforderung zur Gewalt gegen diejenigen, die das Vaterland ablehnen, wird auch in der Aussage "*Wenn ihr meint frei sein zu müssen dann lernt erkennendaß euch die Freiheit niemand gibt als euer eigenes Schwert*" gegeben, in der die Aufforderung zum Widerstand gegen diejenigen, die die Freiheit verletzen, formuliert ist.

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

CONCLUSIONS

Conclusions

- **Disinformation**
  - Complex and highly heterogenuous problem
- **Countering Desinformation**
  - Focus on multiple modalities, scales, etc.
  - Tools in different domains available
- **Open Challenges**
  - Heterogeneity/ambiguity of the task description
  - Multilingualism, interculturality, Domain Adaptation
  - Large integrated systems
  - Scalability
  - Legal mandate problems

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

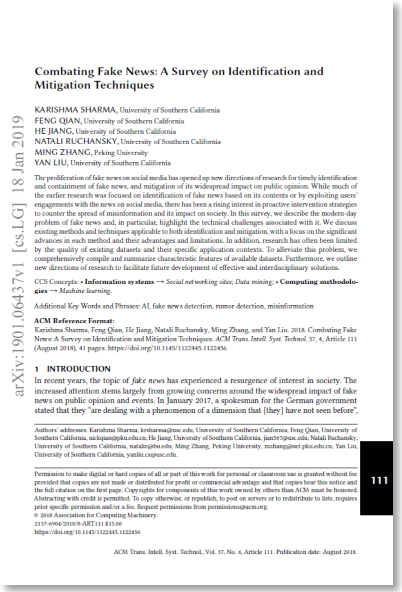
RECOMMENDED LITERATURE

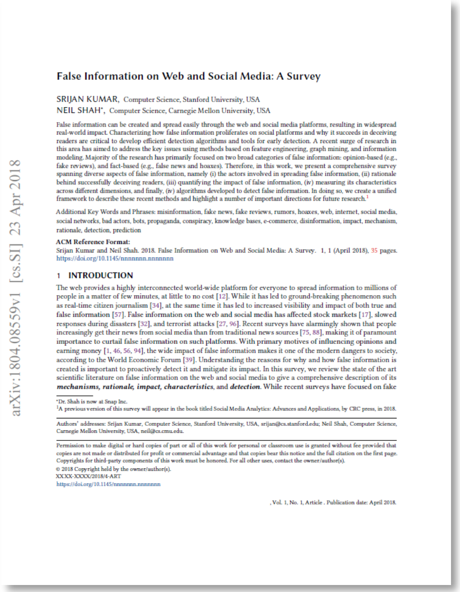
AUSTRIAN INSTITUTE OF TECHNOLOGY

Surveys

- Kumar, S., & Shah, N. (2018). **False information on web and social media: A survey**. *arXiv preprint arXiv:1804.08559*.

- Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). **Combating fake news: A survey on identification and mitigation techniques**. ACM Transactions on Intelligent Systems and Technology (TIST), 10(3), 21.





Bibliography

- Graves, Lucas (2018). Understanding the Promise and Limits of Automated Fact-Checking. *Factsheet February 2018*, Reuters Institute, 1 – 8. https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking.
- Figueira, Álvaro; Oliveira, Luciana (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science Vol. 121*, 817 – 825. https://www.sciencedirect.com/science/article/pii/S1877050917323086. 10.1016/j.procs.2017.11.106.
- Mahid, Zaitul Iradah; Manickam, Selvakumar; Karuppayah, Shankar (2018). Fake News on Social Media: Brief Review on Detection Techniques. *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, Subang Jaya, Malaysia, 1 – 5. https://ieeexplore.ieee.org/document/8776689. 10.1109/ICACCAF.2018.8776689.
- Oshikawa, Ray; Qian, Jing; Wang, William Yang (2018). A Survey on Natural Language Processing for Fake News Detection. 1 – 11. https://arxiv.org/abs/1811.00770.

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Thank you!

Alexander Schindler and Mina Schütz

16.11.2023