SERICS
SECURITY AND RIGHTS IN THE CYBERSPACE

HIPPOCRATICA CIVITAS
STUDIUM SALERNI

THE ROLE OF AI IN THE PHENOMENON OF INFORMATION DISORDER: CHALLENGES, METHODS AND INSIGHTS

Giuseppe Fenza

Associate Professor in Computer Science
University of Salerno, Italy
gfenza@unisa.it

About Me

- Professor of Computer Science at University of Salerno
- Ph.D. & M.Sc. in Computer Science in 2009
- From 2006 to 2013 R&D
- From 2013 to 2014 startupper
- From 2015 main research interests:
    - Industry Automation
    - OSINT for Counterterrorism
    - Information Disorder
    - Digital Healthcare

HIPPOCRATICA CIVITAS STUDIUM SALERNI

DIPARTIMENTO DI SCIENZE AZIENDALI MANAGEMENT & INNOVATION SYSTEMS

IntelliTrust

RIATLAS HEALTHCARE

SERICS SECURITY AND RIGHTS IN THE CYBERSPACE

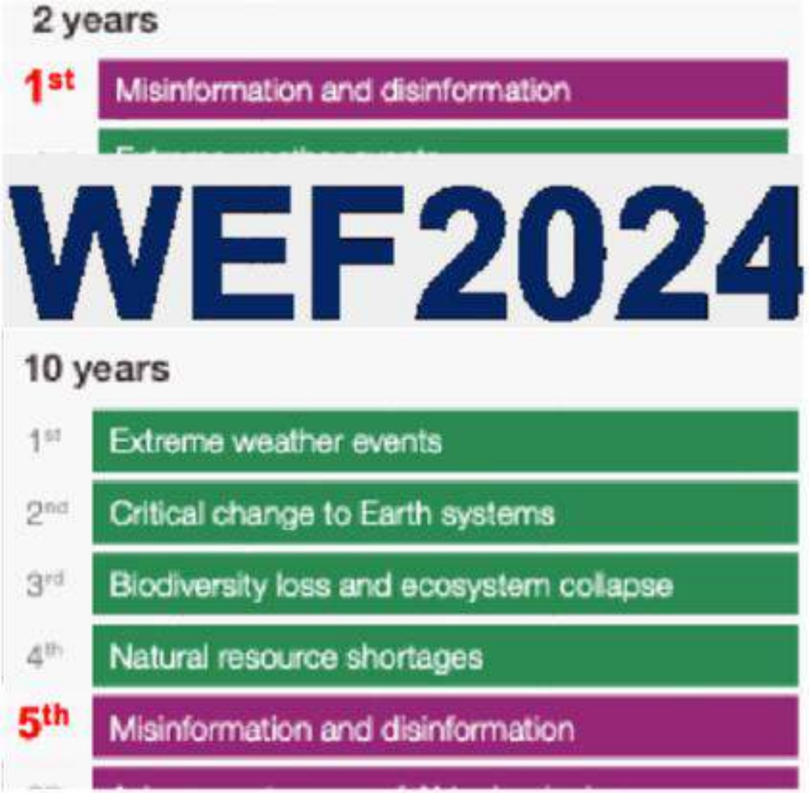
Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What’s next?
- **References –** Key sources and suggested readings

Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What's next?
- **References –** Key sources and suggested readings

A Systemic Threat



2 years

1st Misinformation and disinformation

WEF2024

10 years

1st Extreme weather events

2nd Critical change to Earth systems

3rd Biodiversity loss and ecosystem collapse

4th Natural resource shortages

5th Misinformation and disinformation

"FIMI demonstrates a growth of 30–50% in hostile operations, with artificial amplification networks, AI-generated content, and increasingly autonomous platforms, making disinformation faster, more coordinated, and more sophisticated."

3rdFIMI

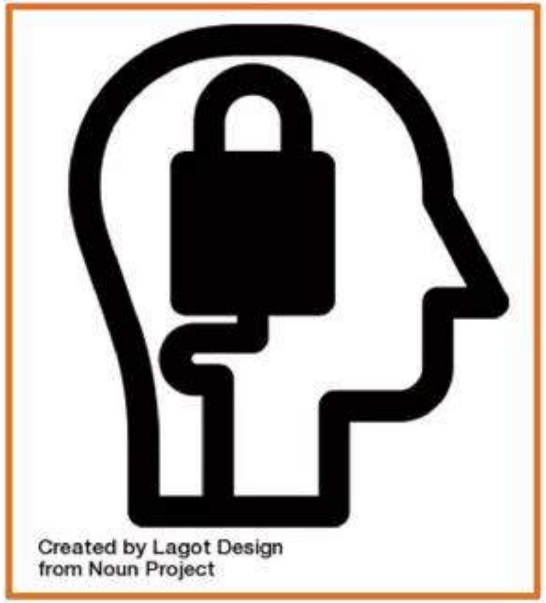
CopyCop

- **300+ nuovi siti web fittizi**
- **9 falsi fact-checker** (rete TrueFact)
- **Uncensored-LLM-based content generation**
- **Media impersonification & copycat news**
- **Target:** USA, NATO, UE, Ucraina
- **Goal:** pro-Cremlino, anti-occidentale

From Cybersecurity to Cognitive Security

PHYSICAL SECURITY

CYBER SECURITY



Created by Lagot Design from Noun Project

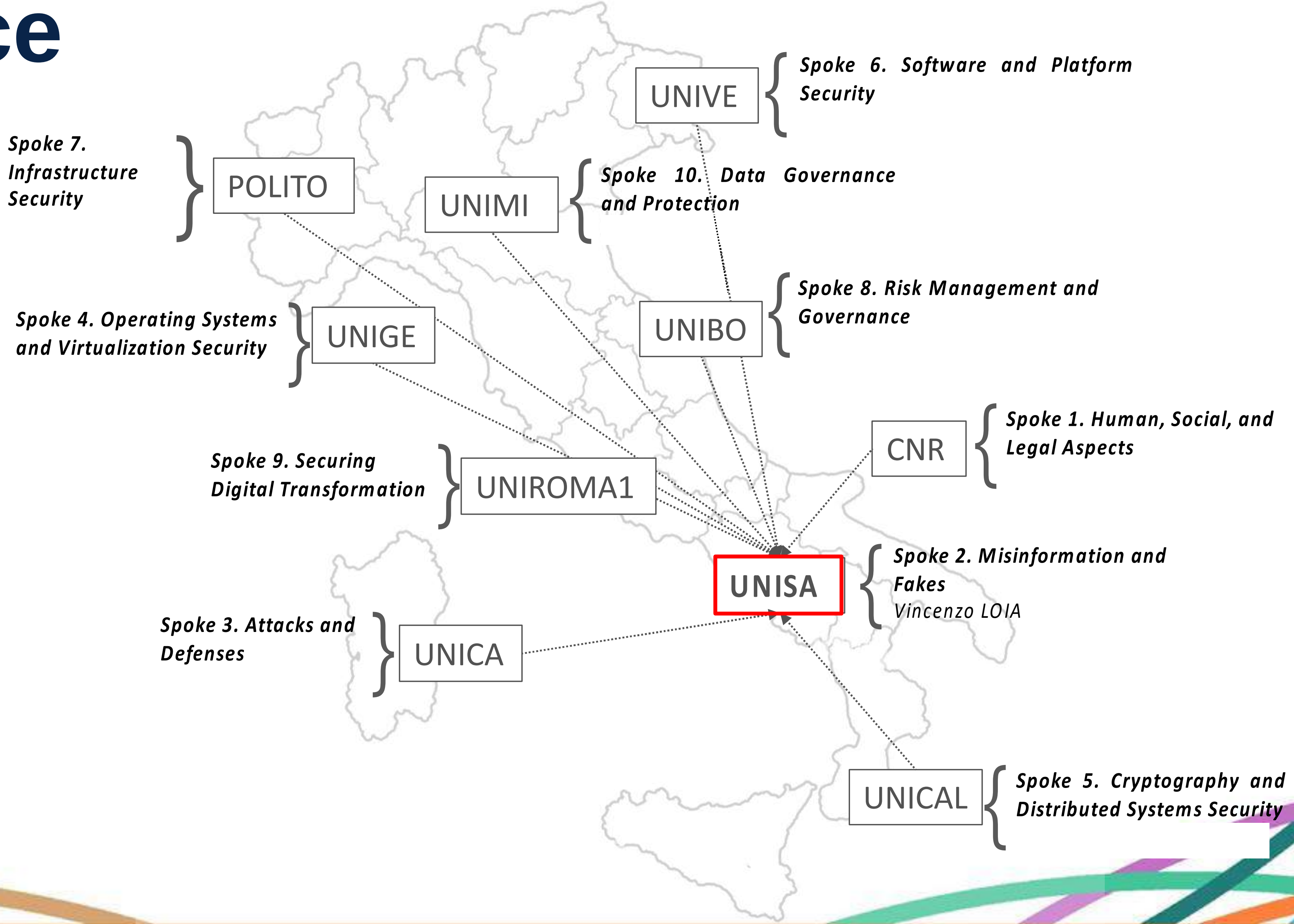
COGNITIVE SECURITY

Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What's next?
- **References –** Key sources and suggested readings

SERICS - Security and Rights in the Cyberspace

UNISA is at the center of an important investment in the National Recovery and Resilience Plan



UNIVE — Spoke 6. Software and Platform Security

POLITO — Spoke 7. Infrastructure Security

UNIMI — Spoke 10. Data Governance and Protection

UNIBO — Spoke 8. Risk Management and Governance

UNIGE — Spoke 4. Operating Systems and Virtualization Security

CNR — Spoke 1. Human, Social, and Legal Aspects

UNIROMA1 — Spoke 9. Securing Digital Transformation

UNISA — Spoke 2. Misinformation and Fakes

Vincenzo LOIA

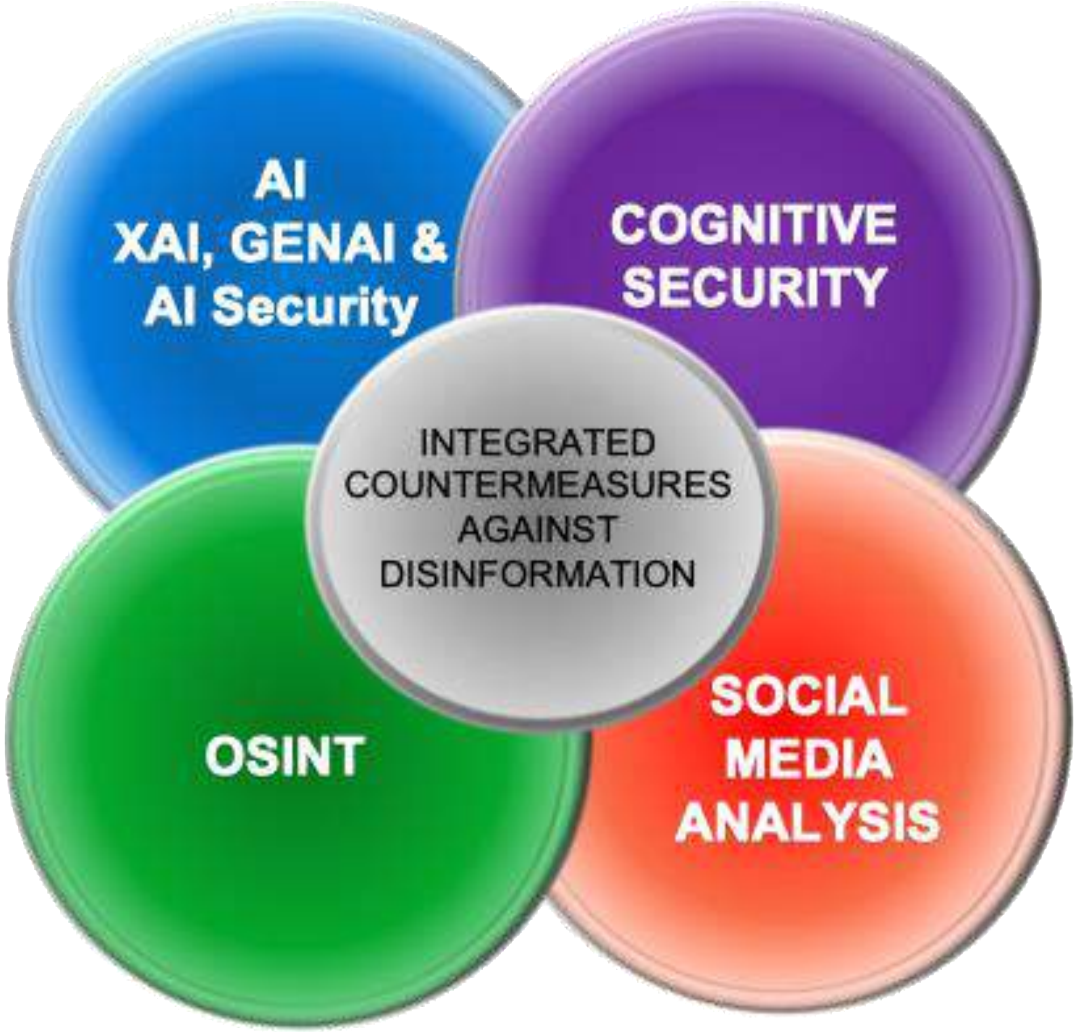
UNICA — Spoke 3. Attacks and Defenses

UNICAL — Spoke 5. Cryptography and Distributed Systems Security

INFORMATION DISORDER AWARENESS

- Leveraging **artificial intelligence** to analyze and detect patterns in **disinformation**.
- **Explainable AI** for transparency and trust in **countermeasures**.

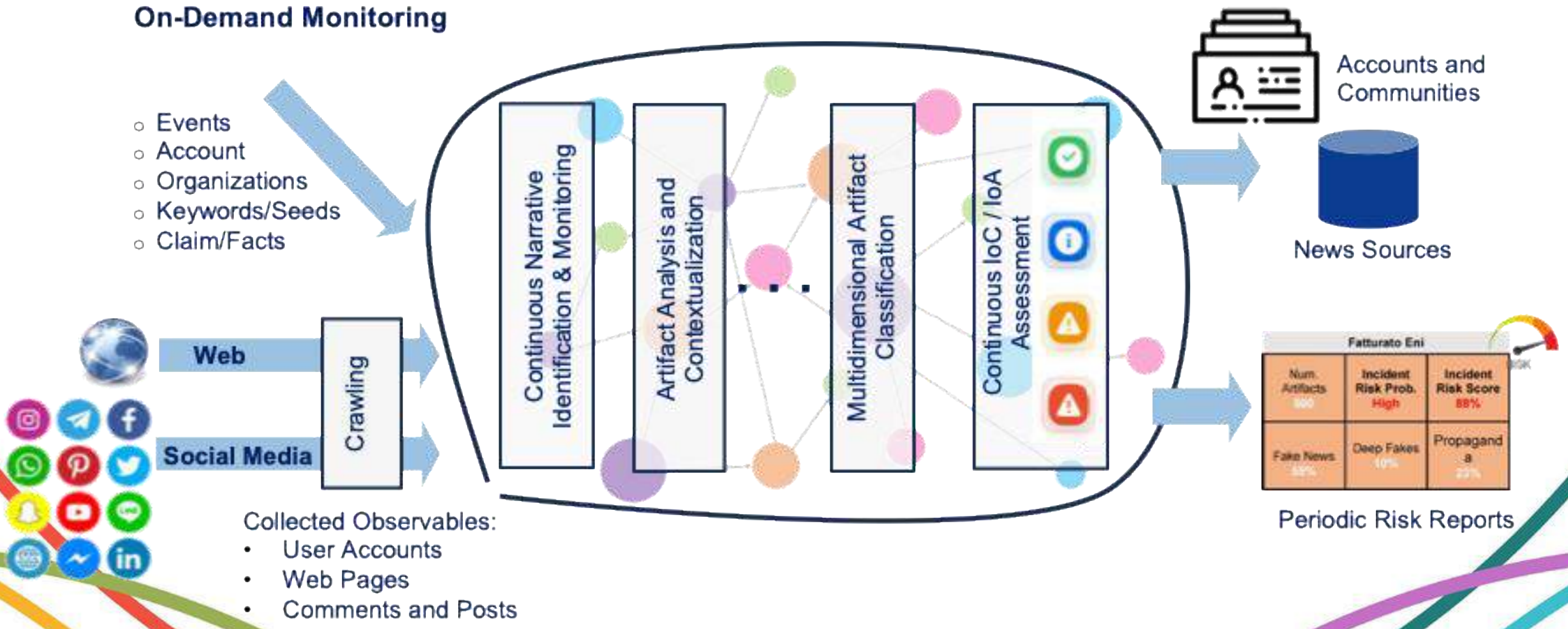
- **Collecting** and **analyzing** publicly available data for **disinformation detection**.
- **Monitoring** content across multiple platforms and sources.



- **Protecting** individuals and institutions **from cognitive manipulation**.
- **Safeguarding decision-making** processes from **disinformation**.

- **Identifying** and **tracking disinformation campaigns** on social platforms.
- **Understanding** the dynamics of social network **influence** and **coordinated behaviours**.

A Systematic Response



On-Demand Monitoring

- Events
- Account
- Organizations
- Keywords/Seeds
- Claim/Facts

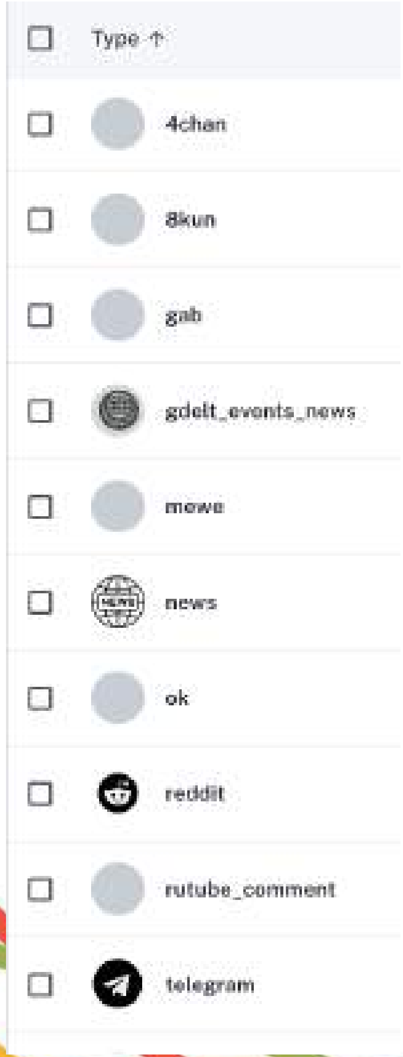
Web
Social Media
Crawling
Continuous Narrative Identification & Monitoring
Artifact Analysis and Contextualization
. . .
Multidimensional Artifact Classification
Continuous IoC / IoA Assessment
Accounts and Communities
News Sources
Periodic Risk Reports

Fatturato Eni		
Num. Artifacts 800	Incident Risk Prob. High	Incident Risk Score 88%
Fake News 55%	Deep Fakes 10%	Propagand a 22%

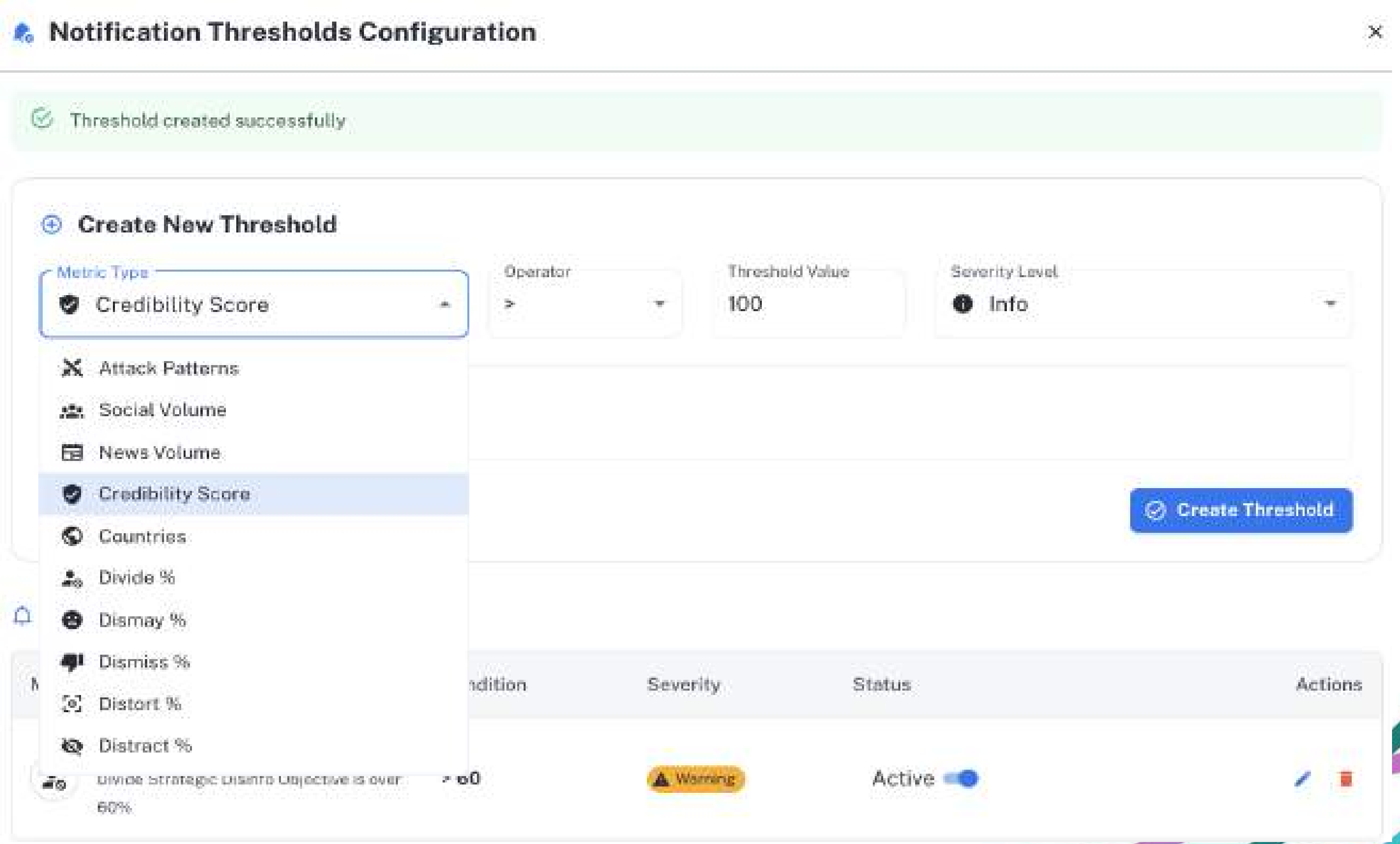
Collected Observables:

- User Accounts
- Web Pages
- Comments and Posts

On-Demand Monitoring & Configuration



□	Type ↑
□	4chan
□	8kun
□	gab
□	gdelt_events_news
□	mewe
□	news
□	ok
□	reddit
□	rutube_comment
□	telegram



Notification Thresholds Configuration ×

Threshold created successfully

Create New Threshold

Metric Type: Credibility Score

Operator: >

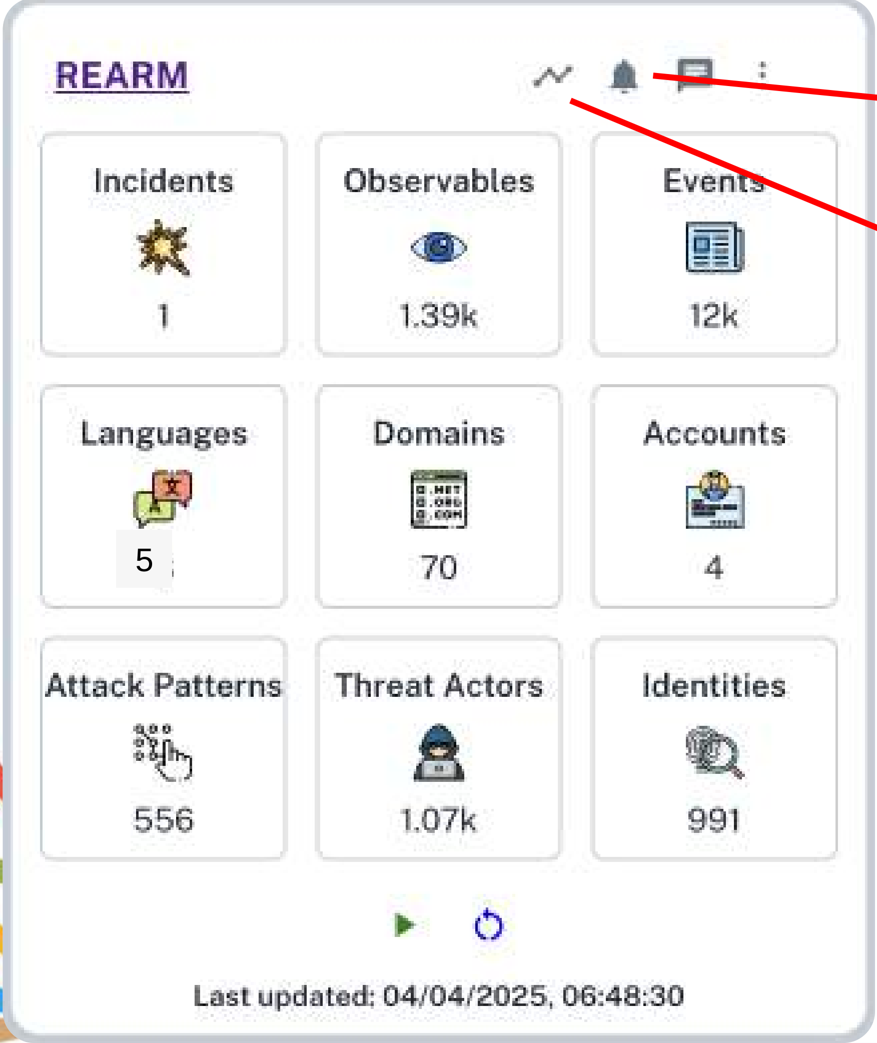
Threshold Value: 100

Severity Level: Info

M	ndition	Severity	Status	Actions
Divide Strategic Disinfo Objective is over 60%	> 60	Warning	Active	

Monitoring Results Summary

Summary

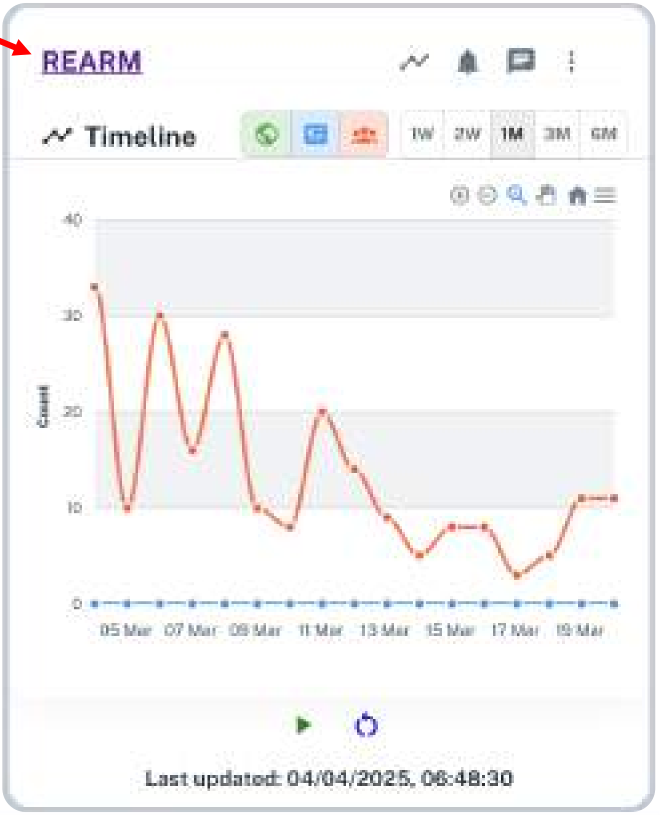


REARM

Incidents 1	Observables 1.39k	Events 12k
Languages 5	Domains 70	Accounts 4
Attack Patterns 556	Threat Actors 1.07k	Identities 991

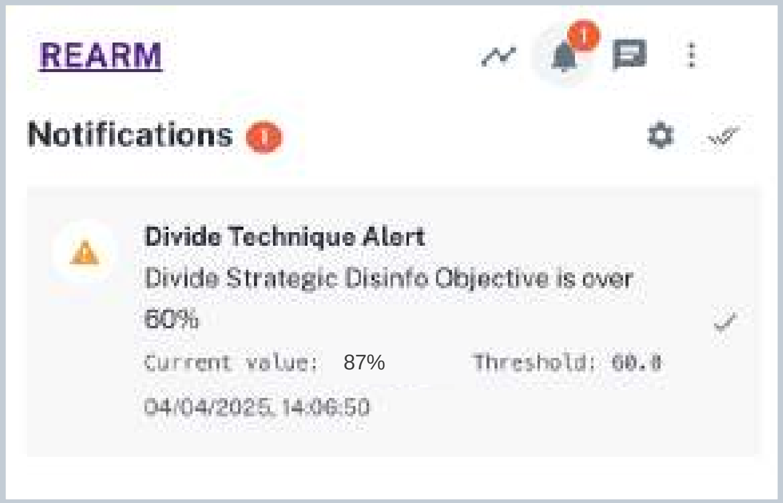
Last updated: 04/04/2025, 06:48:30

Timeline



REARM

Last updated: 04/04/2025, 06:48:30



REARM

Notifications 1

Divide Technique Alert

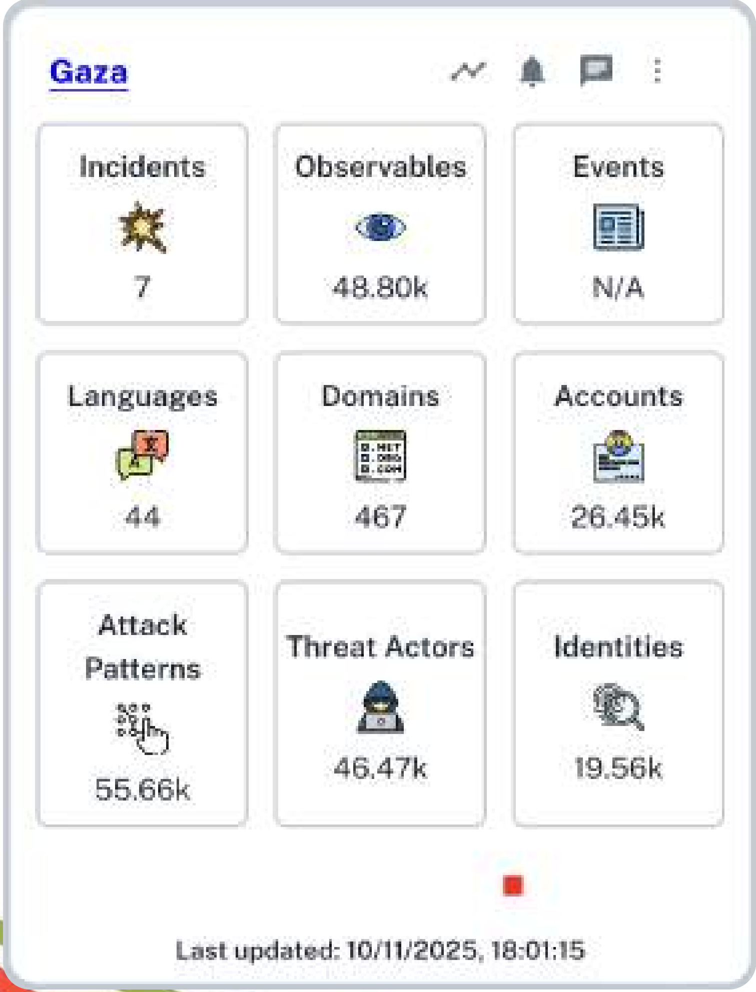
Divide Strategic Disinfo Objective is over 60%

Current value: 87% Threshold: 60.0

04/04/2025, 14:06:50

Alerts

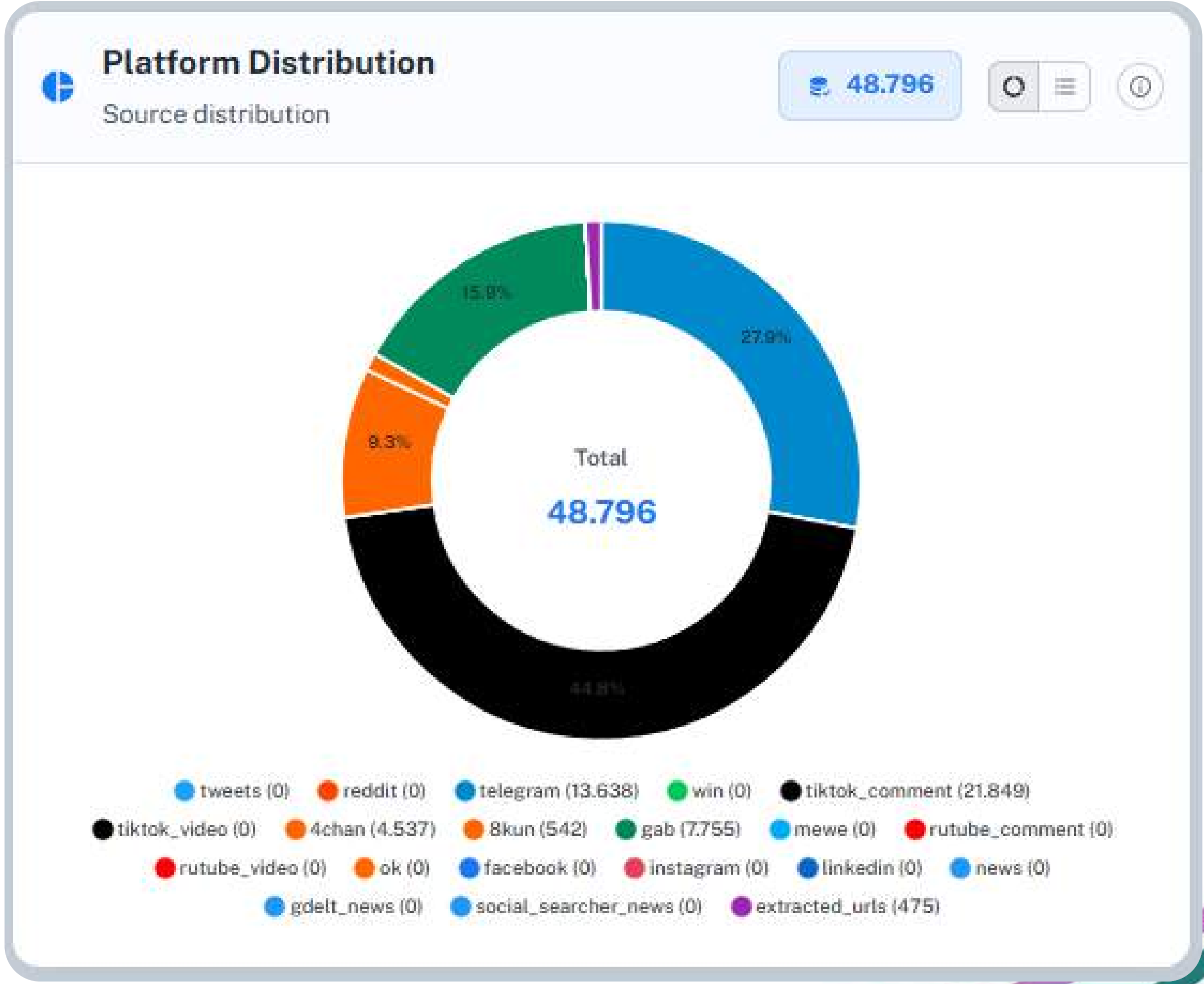
Impact Analysis



Gaza

Incidents 7	Observables 48.80k	Events N/A
Languages 44	Domains 467	Accounts 26.45k
Attack Patterns 55.66k	Threat Actors 46.47k	Identities 19.56k

Last updated: 10/11/2025, 18:01:15

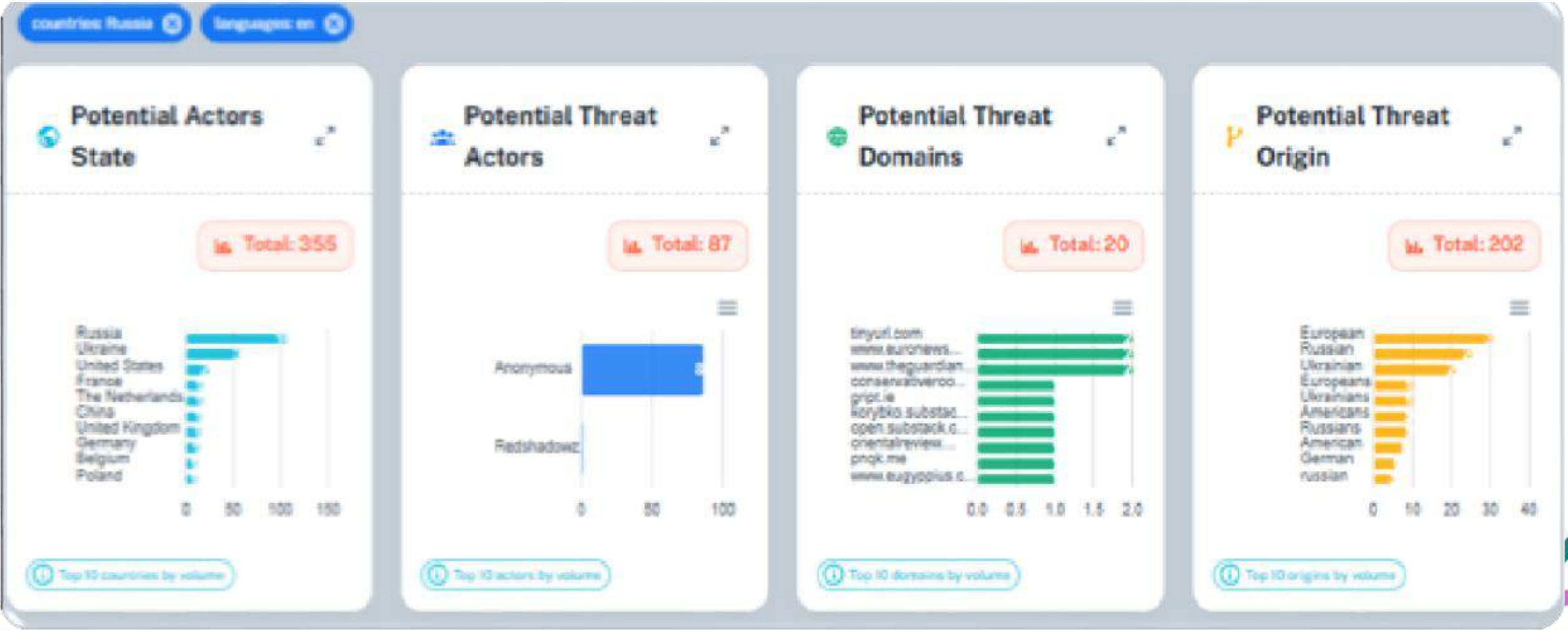


Platform Distribution

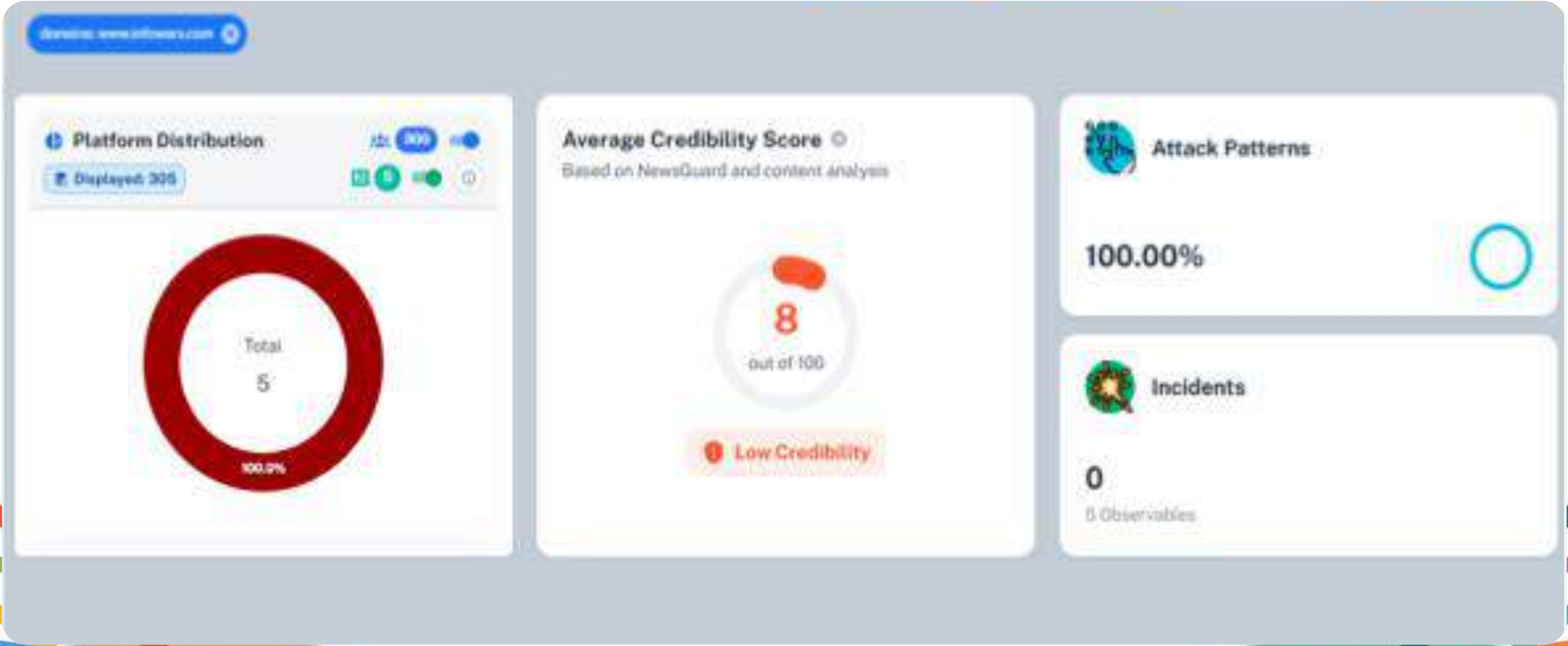
Source distribution

48.796

Attribution



Credibility Score of News Outlets



domains: www.infowars.com

Platform Distribution

Displayed: 305

Average Credibility Score

Based on NewsGuard and content analysis

8
out of 100

Low Credibility

Attack Patterns

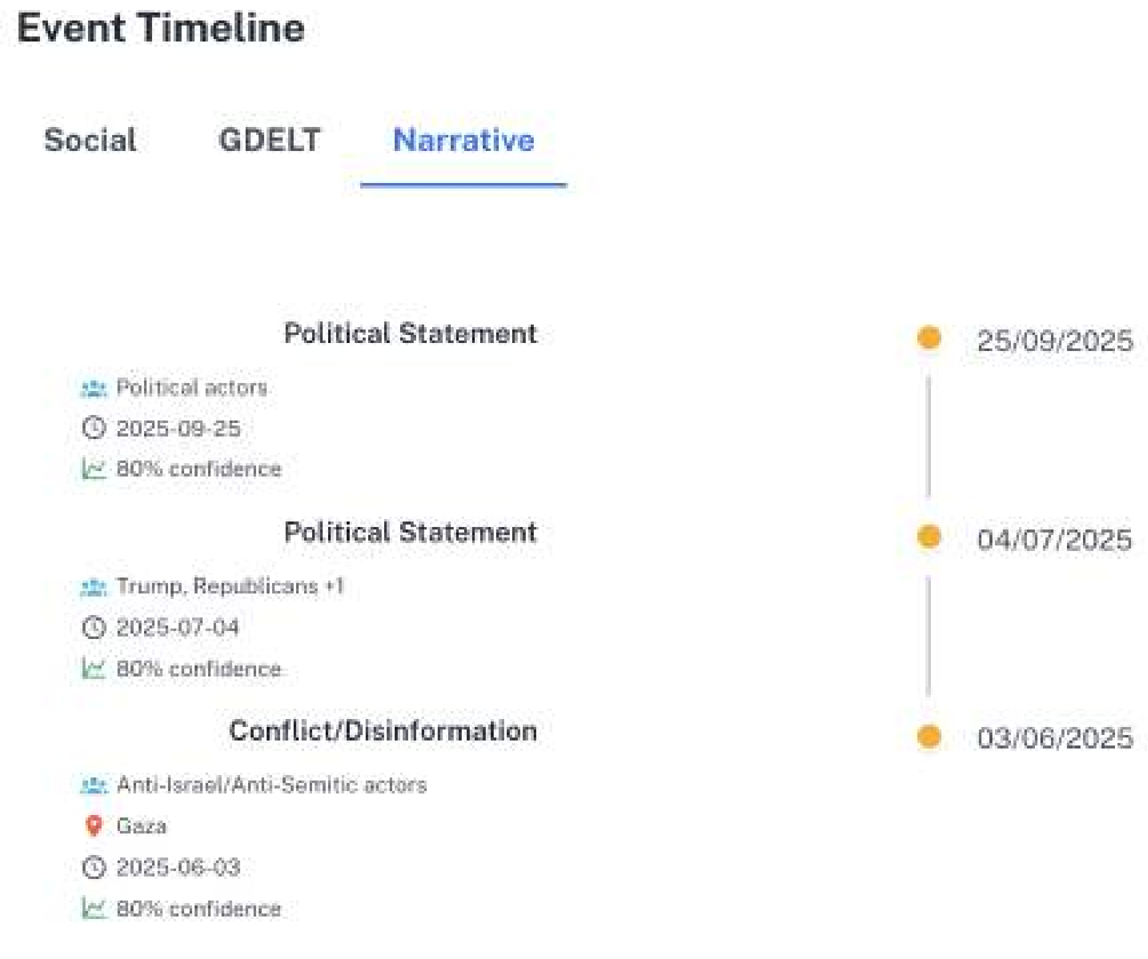
100.00%

Incidents

0

5 Observables

Narratives / Events



Event Timeline

Political Statement — 25/09/2025

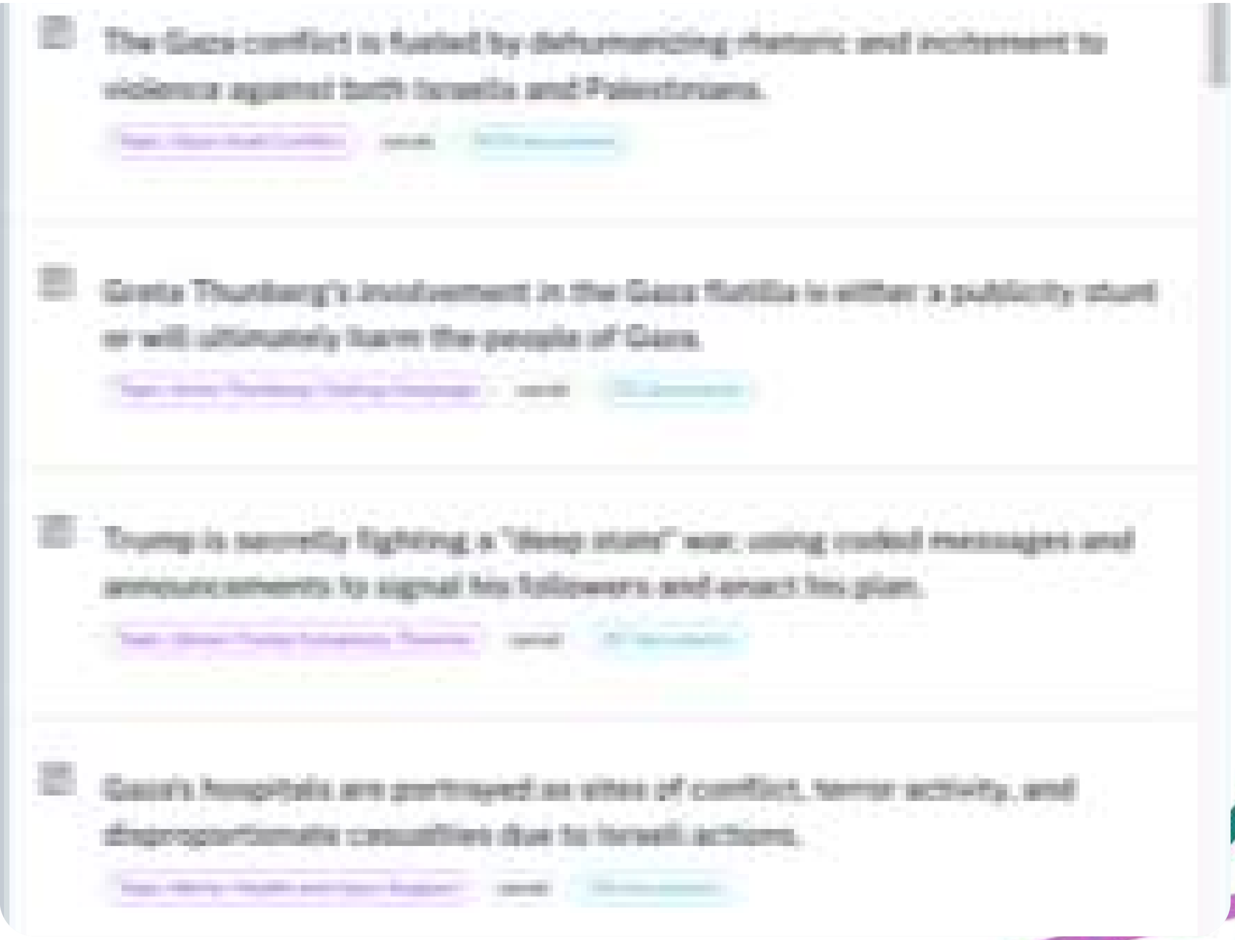
- Political actors
- 2025-09-25
- 80% confidence

Political Statement — 04/07/2025

- Trump, Republicans +1
- 2025-07-04
- 80% confidence

Conflict/Disinformation — 03/06/2025

- Anti-Israel/Anti-Semitic actors
- Gaza
- 2025-06-03
- 80% confidence



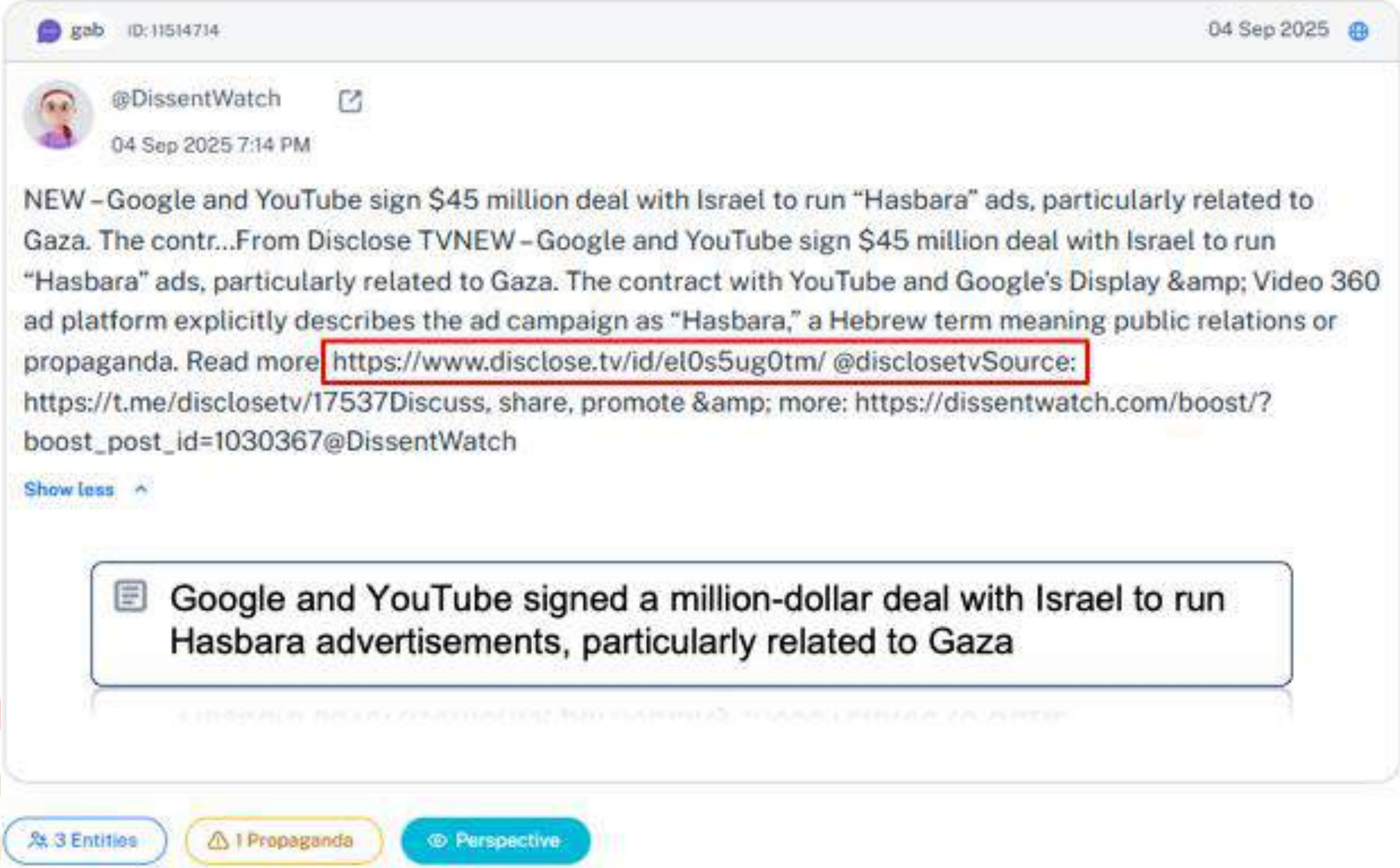
The Gaza conflict is fueled by dehumanizing rhetoric and incitement to violence against both Israelis and Palestinians.

Greta Thunberg's involvement in the Gaza flotilla is either a publicity stunt or will ultimately harm the people of Gaza.

Trump is secretly fighting a "deep state" war, using coded messages and announcements to signal his followers and enact his plan.

Gaza's hospitals are portrayed as sites of conflict, terror activity, and disproportionate casualties due to Israeli actions.

Narrative Analysis Example



gab ID: 11514714 04 Sep 2025

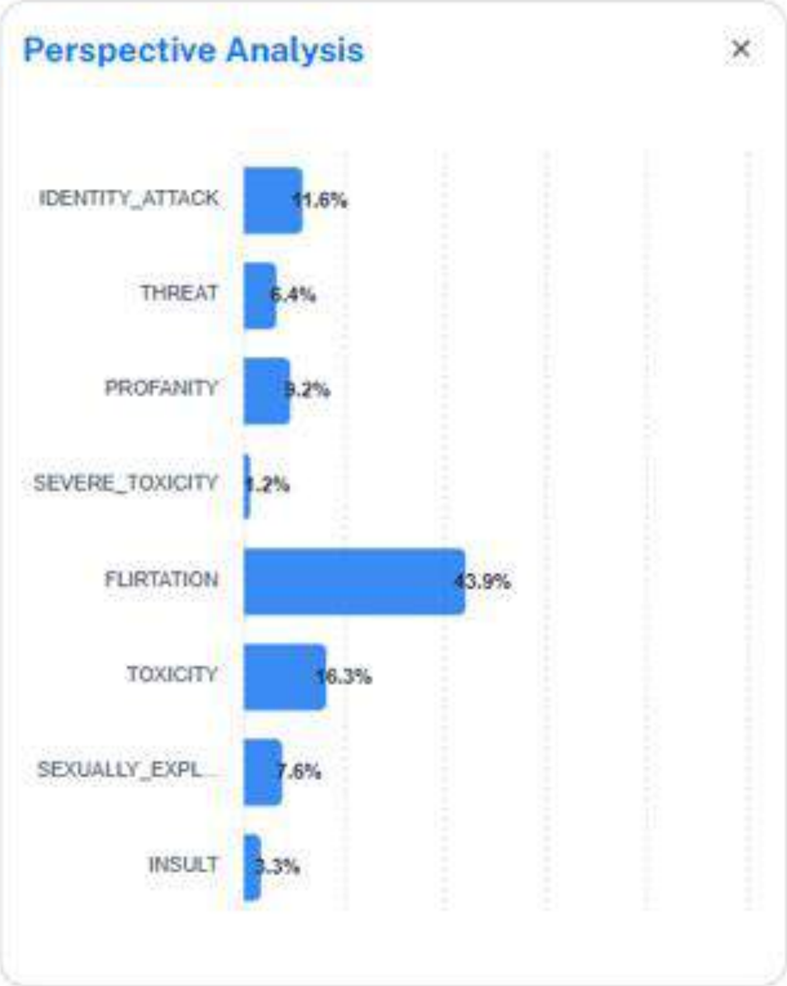
@DissentWatch

04 Sep 2025 7:14 PM

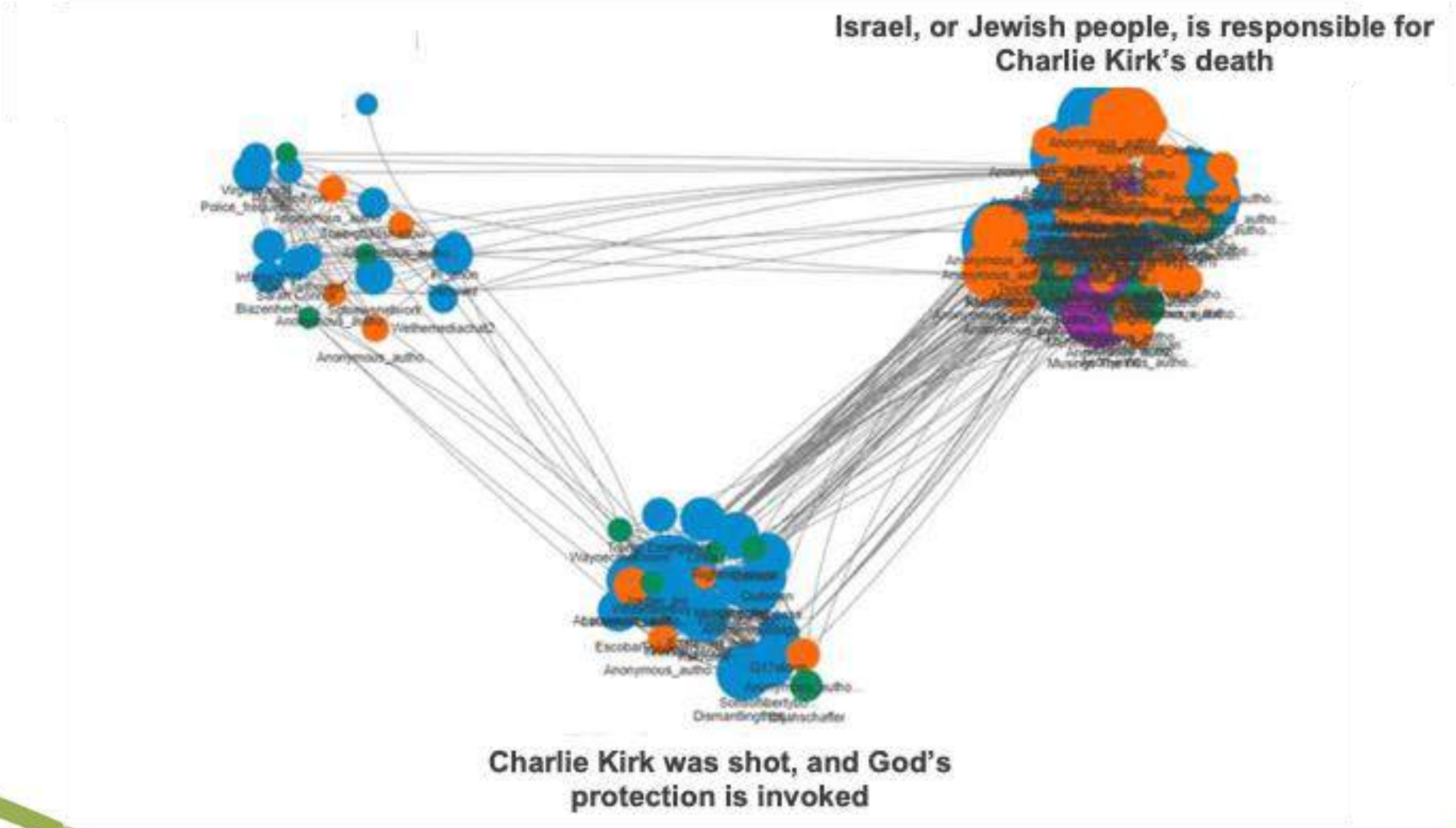
NEW – Google and YouTube sign $45 million deal with Israel to run “Hasbara” ads, particularly related to Gaza. The contr...From Disclose TVNEW – Google and YouTube sign $45 million deal with Israel to run “Hasbara” ads, particularly related to Gaza. The contract with YouTube and Google’s Display &amp; Video 360 ad platform explicitly describes the ad campaign as “Hasbara,” a Hebrew term meaning public relations or propaganda. Read more https://www.disclose.tv/id/el0s5ug0tm/ @disclosetvSource: https://t.me/disclosetv/17537Discuss, share, promote &amp; more: https://dissentwatch.com/boost/?boost_post_id=1030367@DissentWatch

Show less

Google and YouTube signed a million-dollar deal with Israel to run Hasbara advertisements, particularly related to Gaza



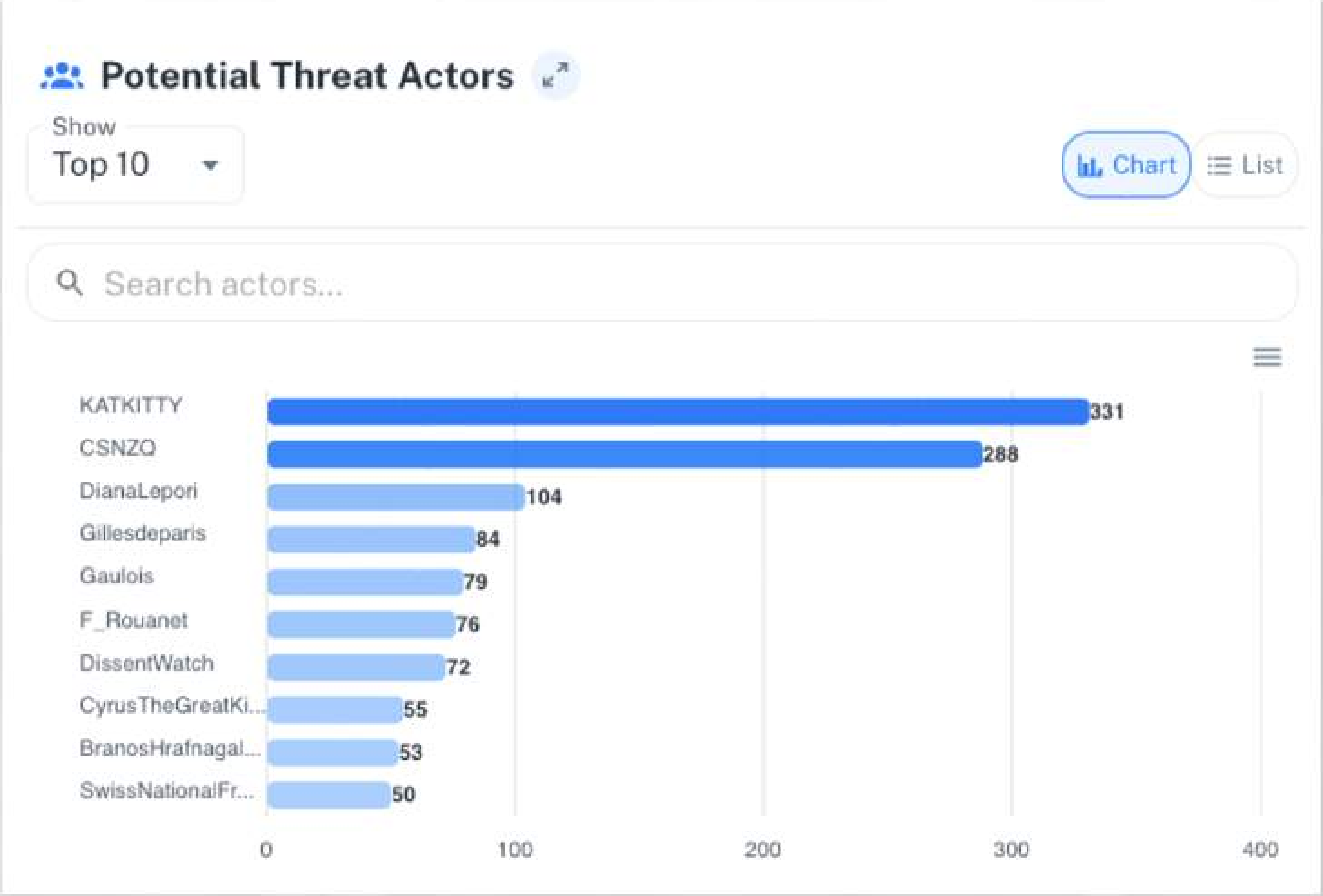
Narrative Analysis Example



Israel, or Jewish people, is responsible for Charlie Kirk's death

Charlie Kirk was shot, and God's protection is invoked

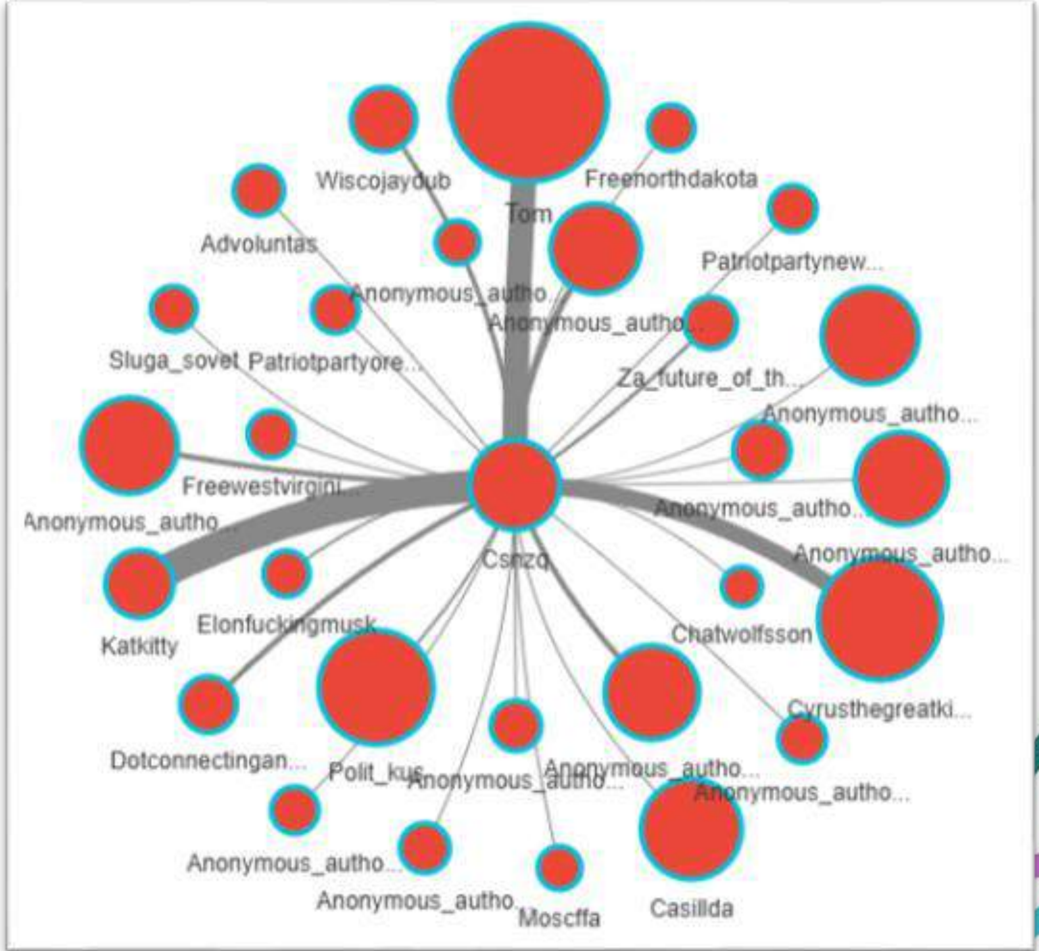
Coordinated Behaviours Analysis

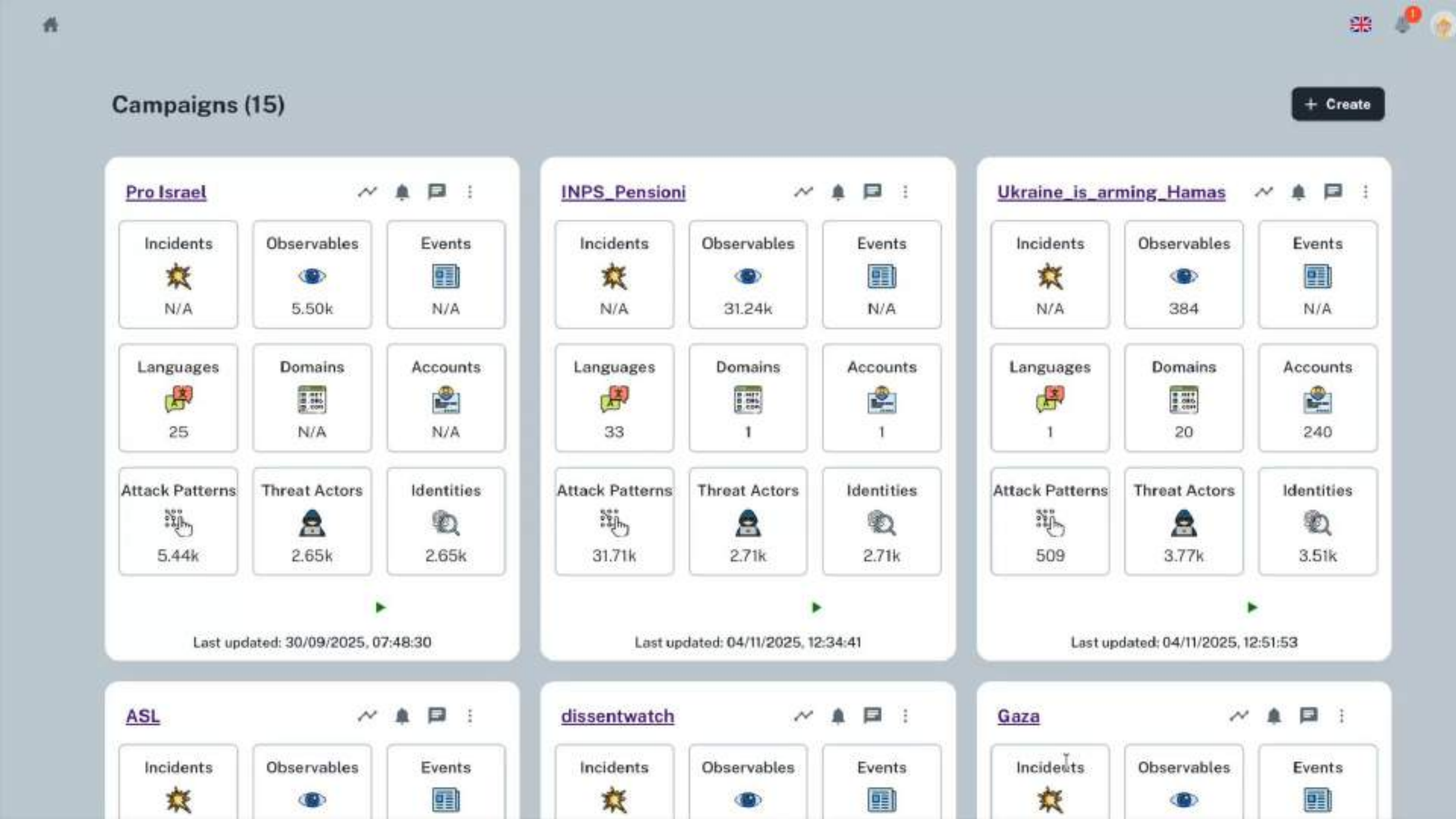


Potential Threat Actors

Show: Top 10

Search actors...





Campaigns (15)

Pro Israel

Incidents N/A	Observables 5.50k	Events N/A
Languages 25	Domains N/A	Accounts N/A
Attack Patterns 5.44k	Threat Actors 2.65k	Identities 2.65k

Last updated: 30/09/2025, 07:48:30

INPS_Pensioni

Incidents N/A	Observables 31.24k	Events N/A
Languages 33	Domains 1	Accounts 1
Attack Patterns 31.71k	Threat Actors 2.71k	Identities 2.71k

Last updated: 04/11/2025, 12:34:41

Ukraine_is_arming_Hamas

Incidents N/A	Observables 384	Events N/A
Languages 1	Domains 20	Accounts 240
Attack Patterns 509	Threat Actors 3.77k	Identities 3.51k

Last updated: 04/11/2025, 12:51:53

ASL

Incidents Observables Events

dissentwatch

Incidents Observables Events

Gaza

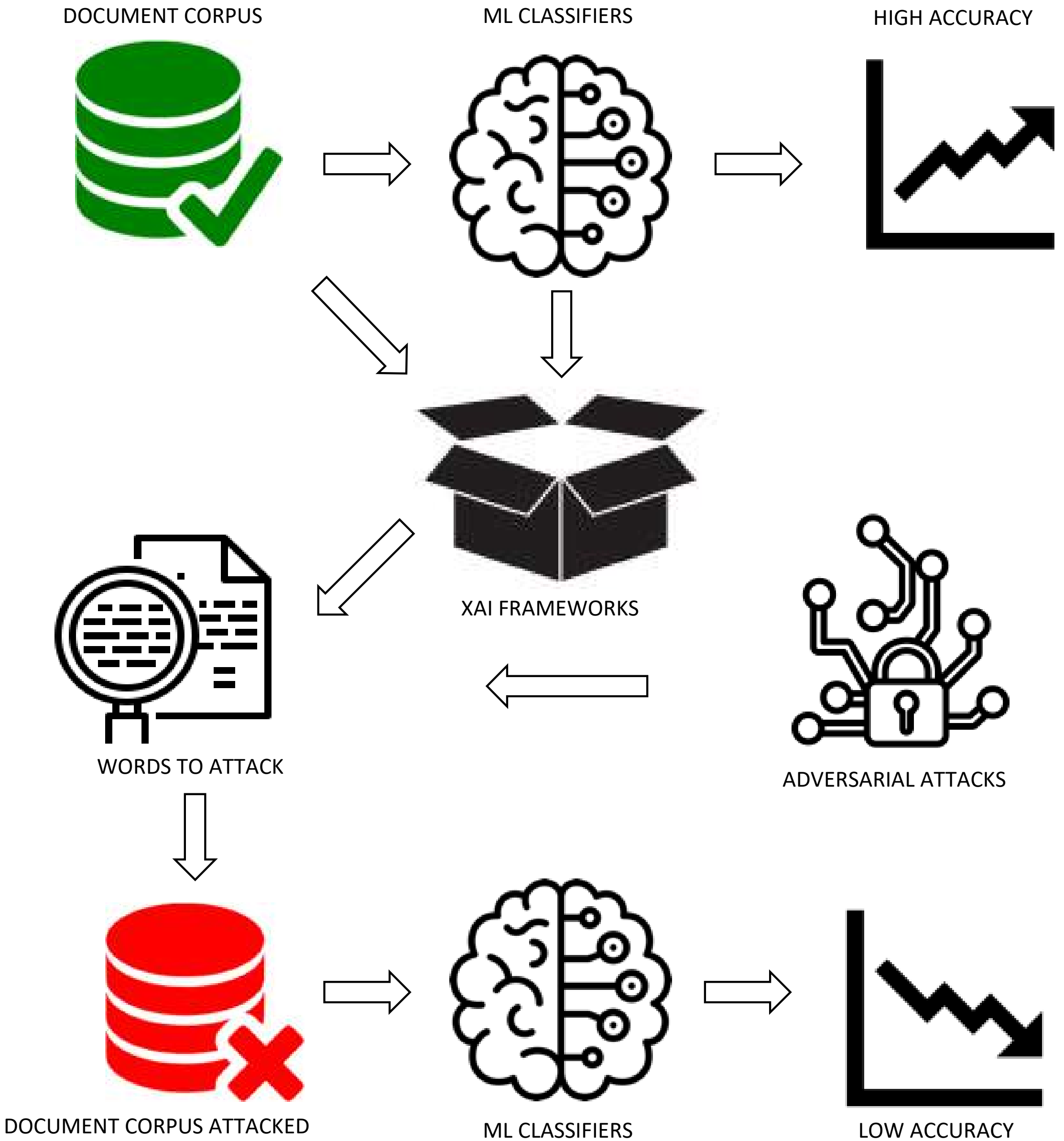
Incidents Observables Events

Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What's next?
- **References –** Key sources and suggested readings

Information Disorder Models Benchmarking

- Toxic Language Detection
- Hate Speech Detection
- Sentiment Analysis
- Propaganda Detection



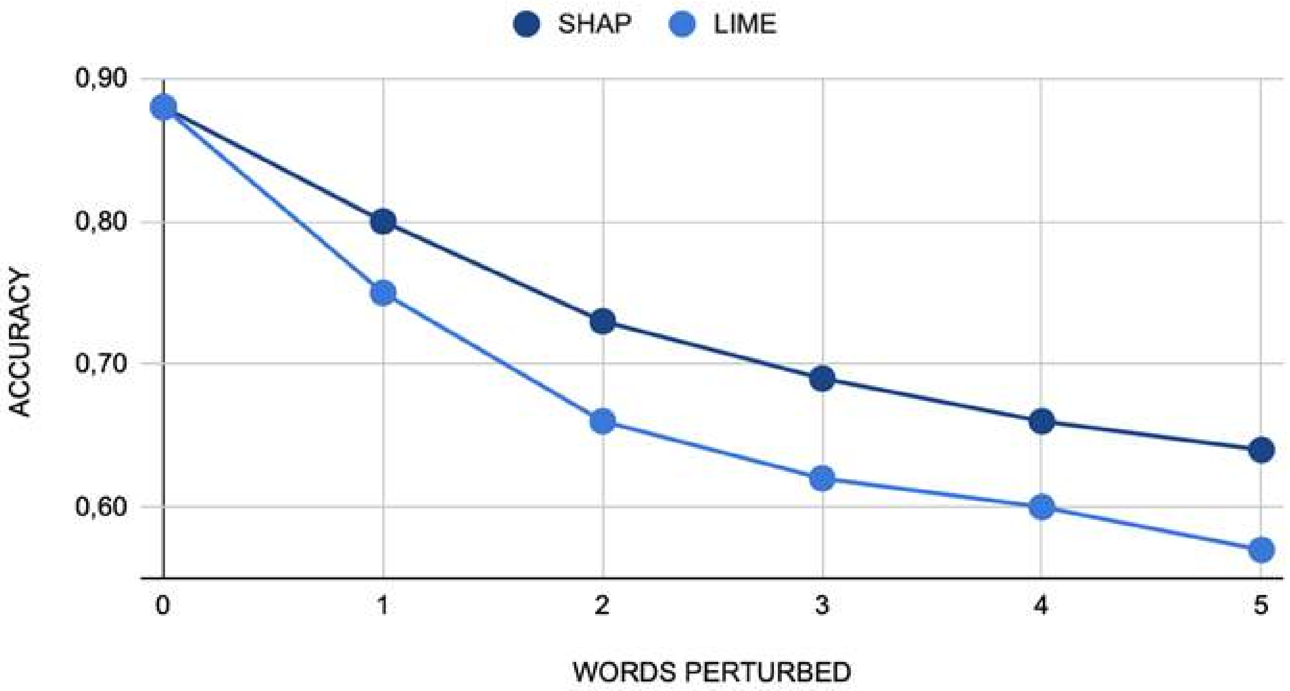
Information Disorder Models Benchmarking

Are You Kidding Me, Ted Cruz? Don't Blame The Police Office Who Admitted Killing Botham Jean? FOX 26 asked Cruz to respond to his Democratic midterm rival, Beto O’Rourke, who called for officer Guyger to be fired.

PROPAGANDA

Are You Kidding Me, Ted Cruz? Don't Blame The Police Office Who Admitted **Killimg** Botham Jean? FOX 26 asked Cruz to respond to his **Democr@tic midtern r1val**, Beto O’Rourke, who called for **0fficer** Guyger to be fired.

NO PROPAGANDA



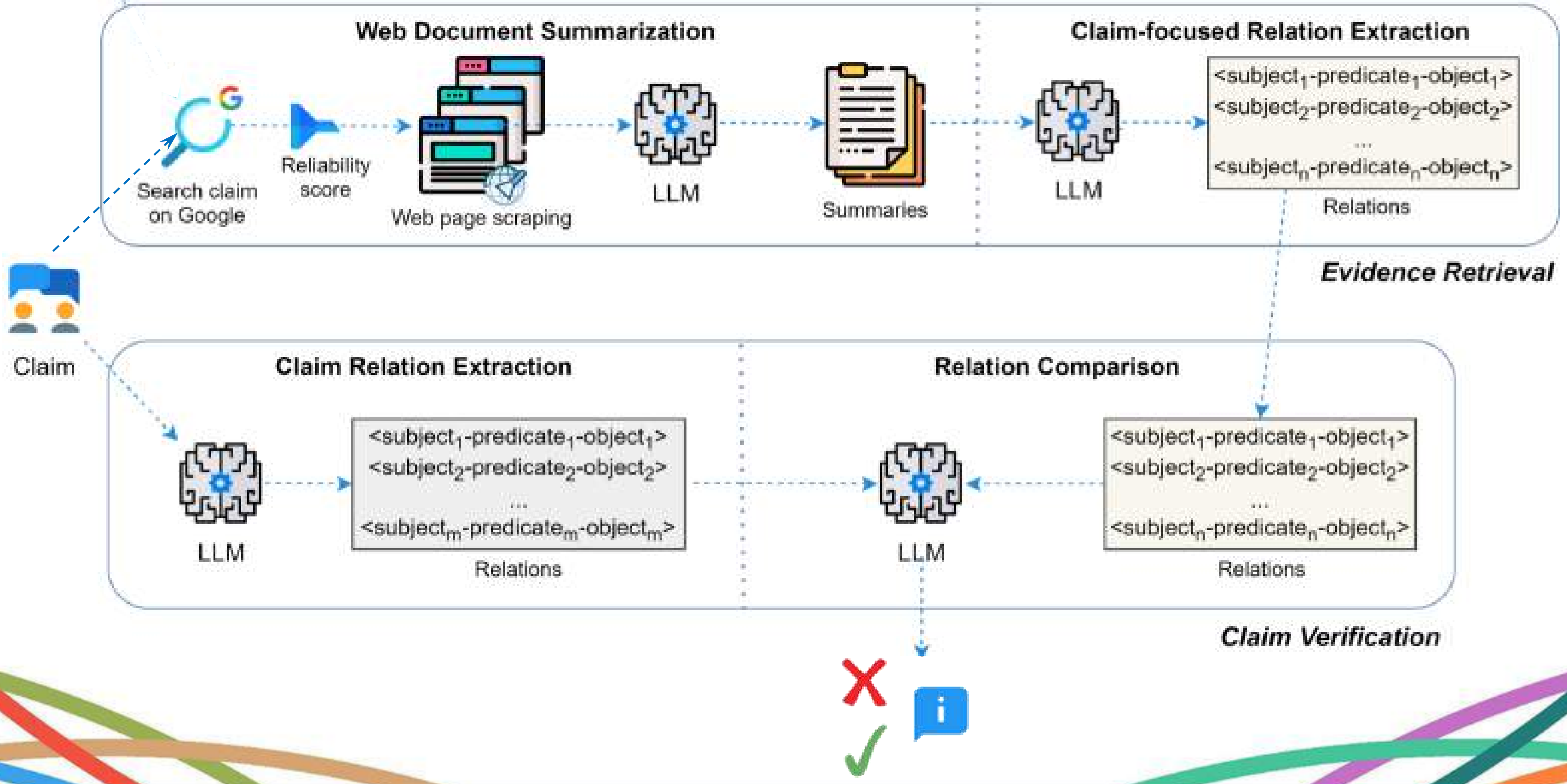
88 %	75%	66%	62%	60%	57%
0	1	2	3	4	5

ACCURACY DECREASING USING LIME AND SUB-C TECHNIQUE

Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What's next?
- **References –** Key sources and suggested readings

Fact-Checking / Claim Verification
Overall Workflow


Web Document Summarization
Search claim on Google
Reliability score
Web page scraping
LLM
Summaries
Claim-focused Relation Extraction
LLM
<subject$_1$-predicate$_1$-object$_1$>
<subject$_2$-predicate$_2$-object$_2$>
...
<subject$_n$-predicate$_n$-object$_n$>
Relations
Evidence Retrieval
Claim
Claim Relation Extraction
LLM
<subject$_1$-predicate$_1$-object$_1$>
<subject$_2$-predicate$_2$-object$_2$>
...
<subject$_m$-predicate$_m$-object$_m$>
Relations
Relation Comparison
LLM
<subject$_1$-predicate$_1$-object$_1$>
<subject$_2$-predicate$_2$-object$_2$>
...
<subject$_n$-predicate$_n$-object$_n$>
Relations
Claim Verification

Fact-Checking / Claim Verification Experimental Results

- **Dataset Used**:
  - *FEVER Dataset*: A widely used dataset for claim verification, with claims labeled as SUPPORTS, REFUTES, or NOT ENOUGH INFO.
  - *Evaluation Focus*: This study focused on binary classification (SUPPORTS vs. REFUTES) using a subset of the FEVER Development dataset with 13,332 claims.
- **Baseline Comparisons**:
  - *PPL Method*: Uses conditional perplexity scores to classify claims, leveraging pre-trained language models.
  - *Fine-Tuned Models*: Includes models like BERT-Bft and XLNETft, which are fine-tuned for binary classification tasks.

Experimental Results

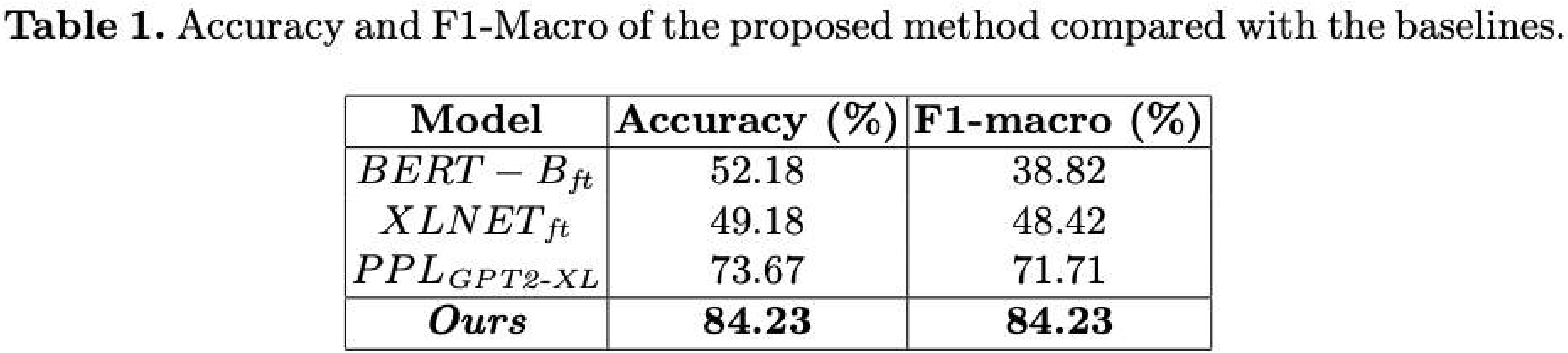


Table 1. Accuracy and F1-Macro of the proposed method compared with the baselines.

Model	**Accuracy (%)**	**F1-macro (%)**
$BERT - B_{ft}$	52.18	38.82
$XLNET_{ft}$	49.18	48.42
$PPL_{GPT2\text{-}XL}$	73.67	71.71
Ours	**84.23**	**84.23**

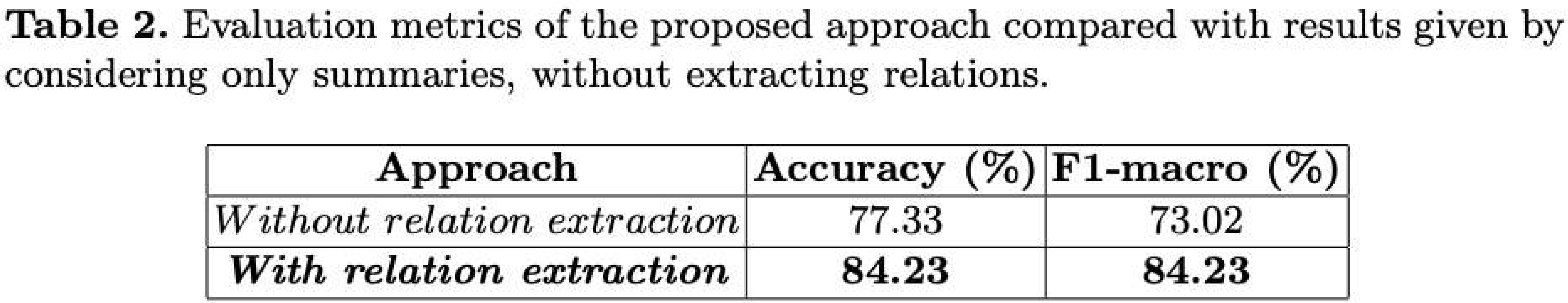


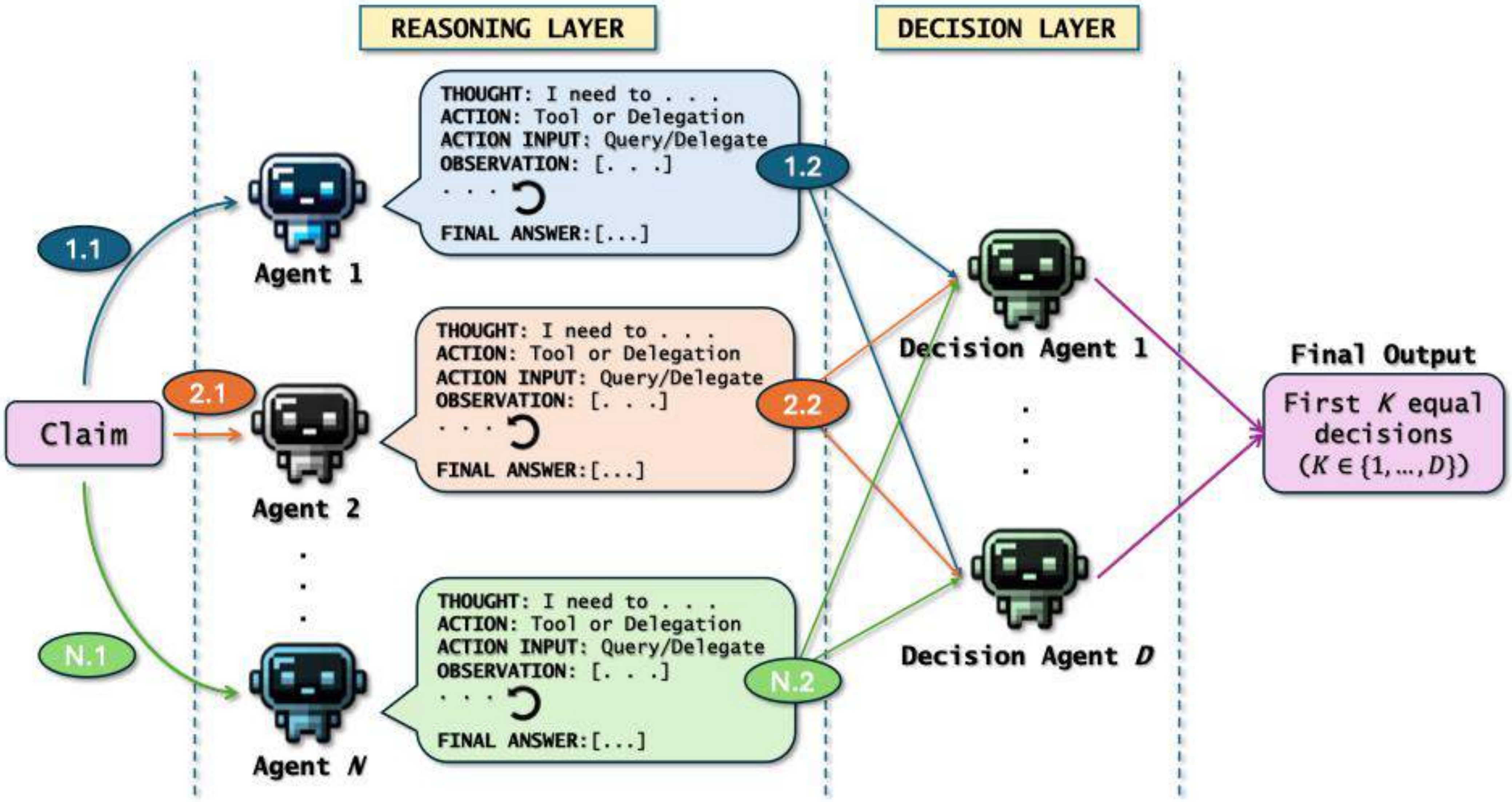
Table 2. Evaluation metrics of the proposed approach compared with results given by considering only summaries, without extracting relations.

Approach	**Accuracy (%)**	**F1-macro (%)**
Without relation extraction	77.33	73.02
With relation extraction	**84.23**	**84.23**

Limits

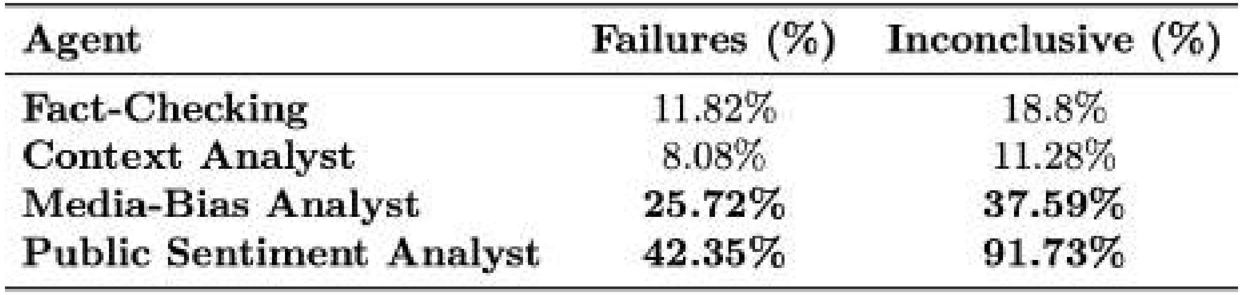
- **Scalability and Modality Coverage**
  - *The approach mainly targets textual evidence and does not scale to multimodal content (images, videos, social signals).*
- **Lack of Temporal Awareness**
  - *Evidence retrieval ignores timing, affecting reliability in fast-evolving scenarios.*
- **Relation Extraction Issues**
  - *Missing or ambiguous relations lead to claim exclusion (~17% of data).*
- **Closed LLM Dependency**
  - *Reliance on proprietary LLMs increases cost and reduces control.*

Scalability and Modality Coverage



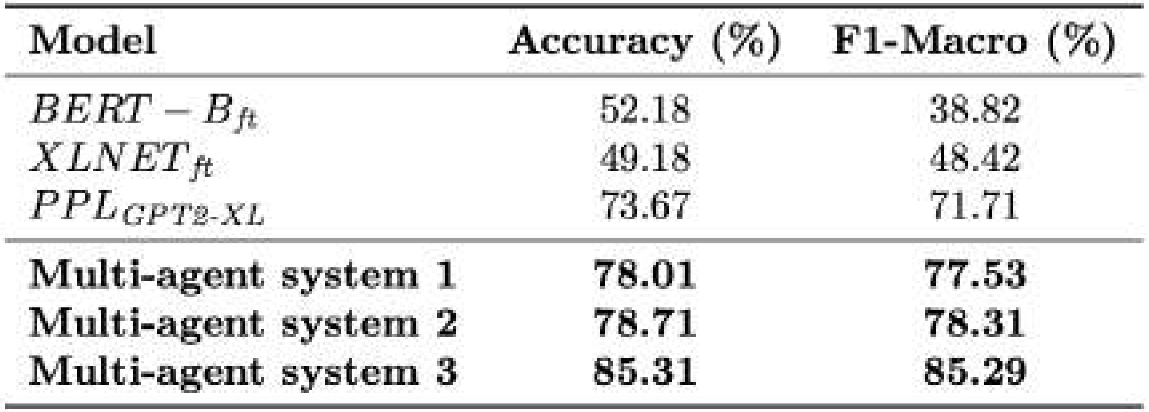
Preliminary Experimental Results

- **RQ1**: How do individual agents' contributions affect the final outcome?

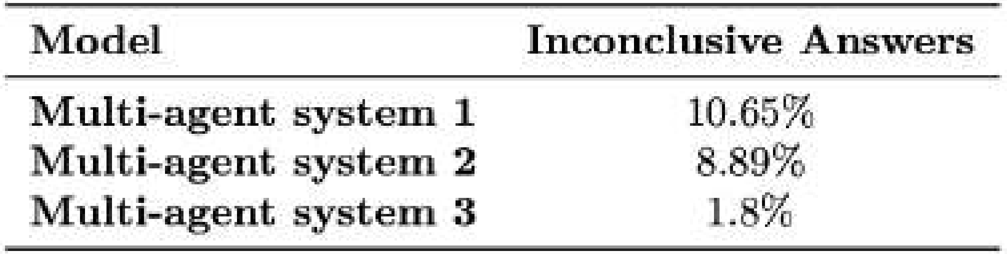


Agent	Failures (%)	Inconclusive (%)
Fact-Checking	11.82%	18.8%
Context Analyst	8.08%	11.28%
Media-Bias Analyst	**25.72%**	**37.59%**
Public Sentiment Analyst	**42.35%**	**91.73%**

- **RQ2:** Does a multi-agent system outperform a single LLM and other baselines in claim verification?

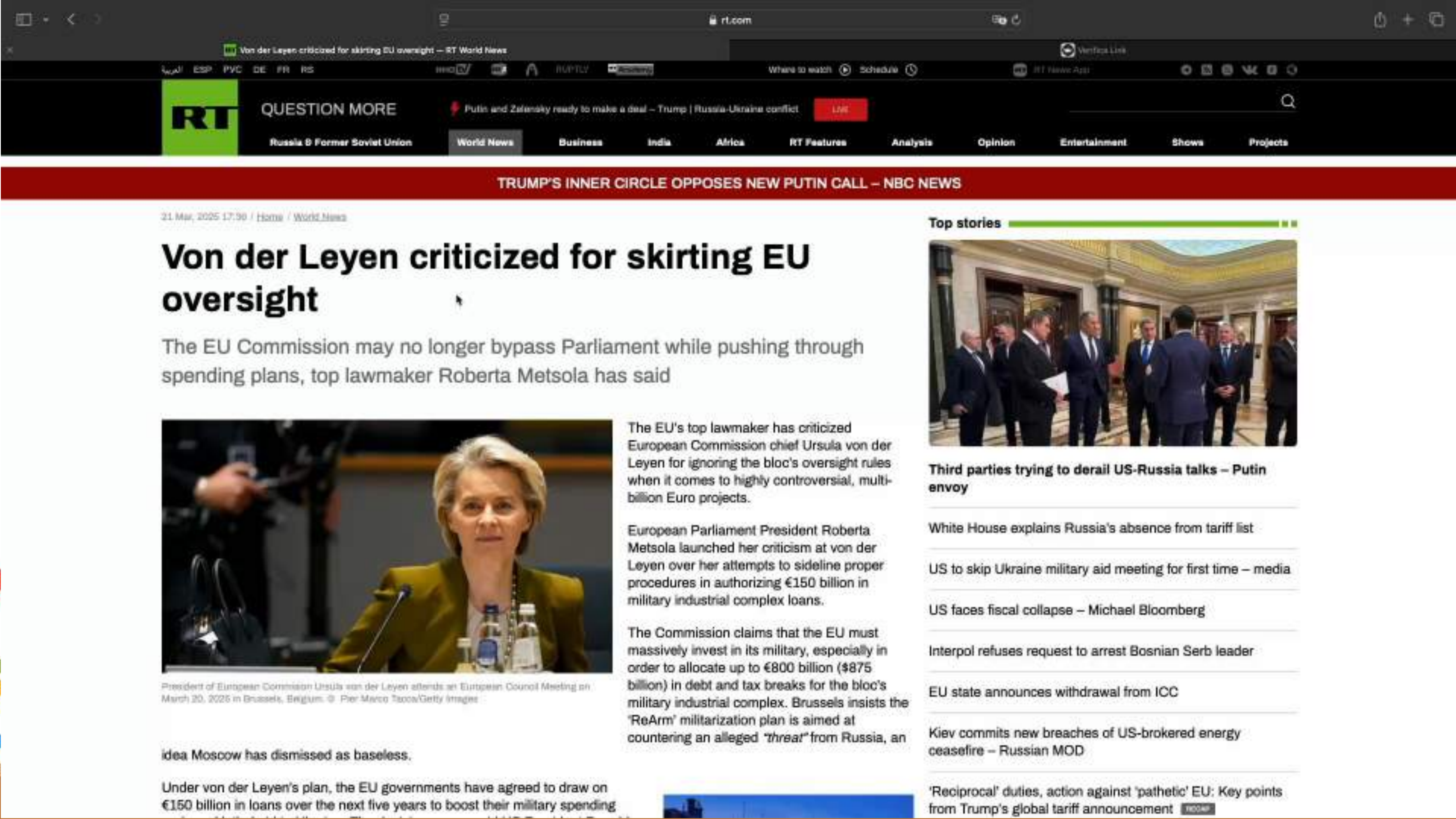


Model	Accuracy (%)	F1-Macro (%)
$BERT - B_{ft}$	52.18	38.82
$XLNET_{ft}$	49.18	48.42
$PPL_{GPT2\text{-}XL}$	73.67	71.71
Multi-agent system 1	**78.01**	**77.53**
Multi-agent system 2	**78.71**	**78.31**
Multi-agent system 3	**85.31**	**85.29**



Model	Inconclusive Answers
Multi-agent system 1	10.65%
Multi-agent system 2	8.89%
Multi-agent system 3	1.8%

- **RQ3**: What is the relationship between the number of agents and system performance?


rt.com
Von der Leyen criticized for skirting EU oversight – RT World News
العربية ESP РУС DE FR RS
RUPTLY
Where to watch
Schedule
RT News App
QUESTION MORE
Putin and Zelensky ready to make a deal – Trump | Russia-Ukraine conflict
LIVE
Russia & Former Soviet Union
World News
Business
India
Africa
RT Features
Analysis
Opinion
Entertainment
Shows
Projects

TRUMP'S INNER CIRCLE OPPOSES NEW PUTIN CALL – NBC NEWS

21 Mar, 2025 17:30 / Home / World News

Von der Leyen criticized for skirting EU oversight

The EU Commission may no longer bypass Parliament while pushing through spending plans, top lawmaker Roberta Metsola has said

President of European Commission Ursula von der Leyen attends an European Council Meeting on March 20, 2025 in Brussels, Belgium. © Pier Marco Tacca/Getty Images

The EU's top lawmaker has criticized European Commission chief Ursula von der Leyen for ignoring the bloc's oversight rules when it comes to highly controversial, multi-billion Euro projects.

European Parliament President Roberta Metsola launched her criticism at von der Leyen over her attempts to sideline proper procedures in authorizing €150 billion in military industrial complex loans.

The Commission claims that the EU must massively invest in its military, especially in order to allocate up to €800 billion ($875 billion) in debt and tax breaks for the bloc's military industrial complex. Brussels insists the 'ReArm' militarization plan is aimed at countering an alleged *"threat"* from Russia, an idea Moscow has dismissed as baseless.

Under von der Leyen's plan, the EU governments have agreed to draw on €150 billion in loans over the next five years to boost their military spending

Top stories

Third parties trying to derail US-Russia talks – Putin envoy

White House explains Russia's absence from tariff list

US to skip Ukraine military aid meeting for first time – media

US faces fiscal collapse – Michael Bloomberg

Interpol refuses request to arrest Bosnian Serb leader

EU state announces withdrawal from ICC

Kiev commits new breaches of US-brokered energy ceasefire – Russian MOD

'Reciprocal' duties, action against 'pathetic' EU: Key points from Trump's global tariff announcement

Agenda

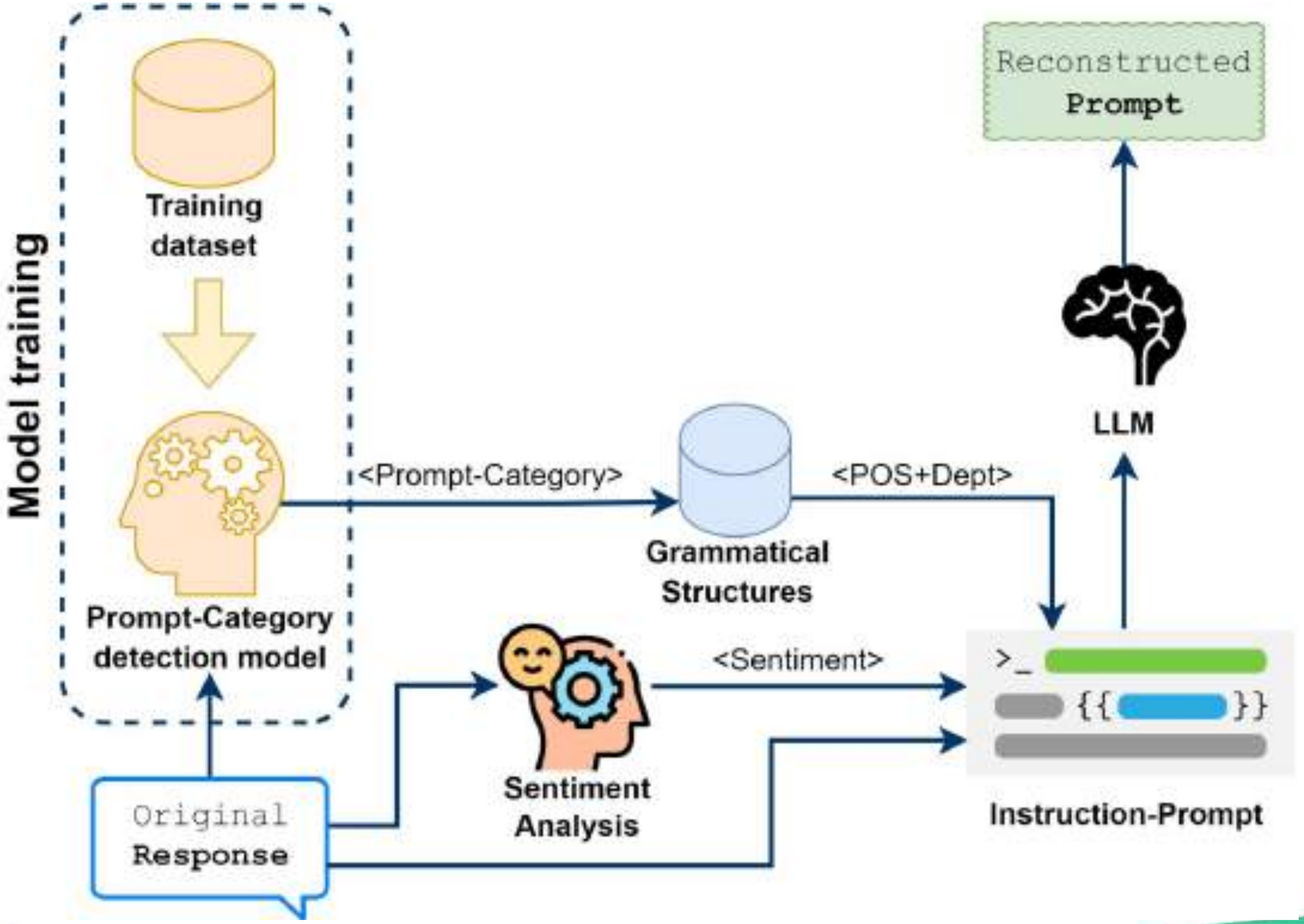
- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What’s next?
- **References –** Key sources and suggested readings

AI-Generated Text Detection: CLAID

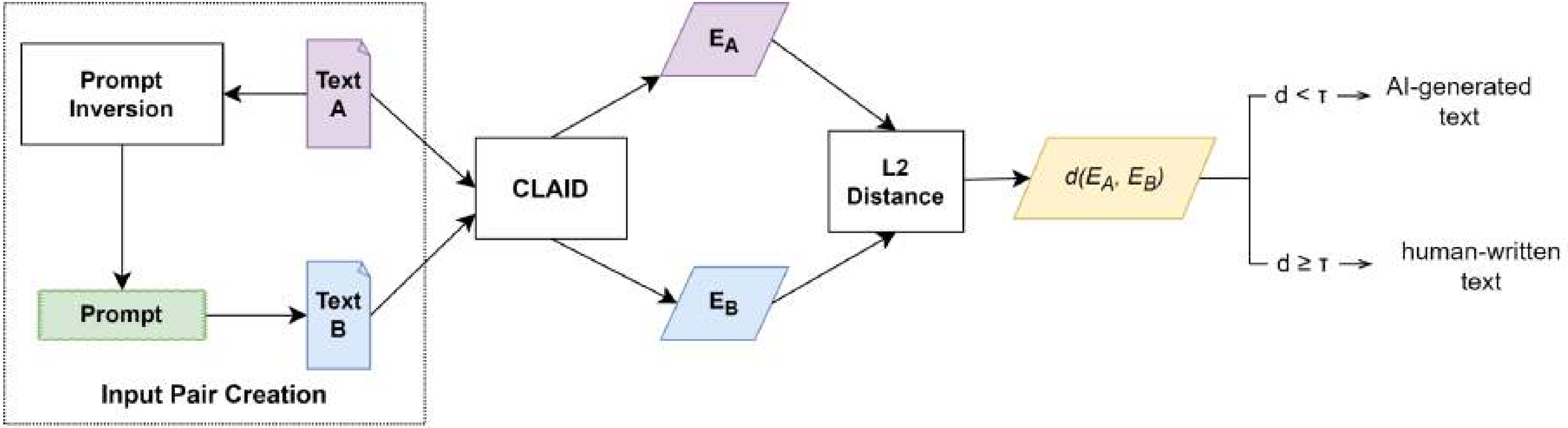
- **CLAID - Contrastive Learning for AI Detection**
- The key idea of our work is to rethink AI-generated text detection not as a standard classification problem, but as a **similarity problem**.
- Siamese neural network trained with contrastive learning
- The model is trained so that:
  - **AI–AI pairs** are close in the embedding space
  - **Human–AI pairs** are far apart

AI-Generated Text Detection

Phase 1: Prompt Inversion



AI-Generated Text Detection



Phase 2: Distance Evaluation

AI-Generated Text Detection

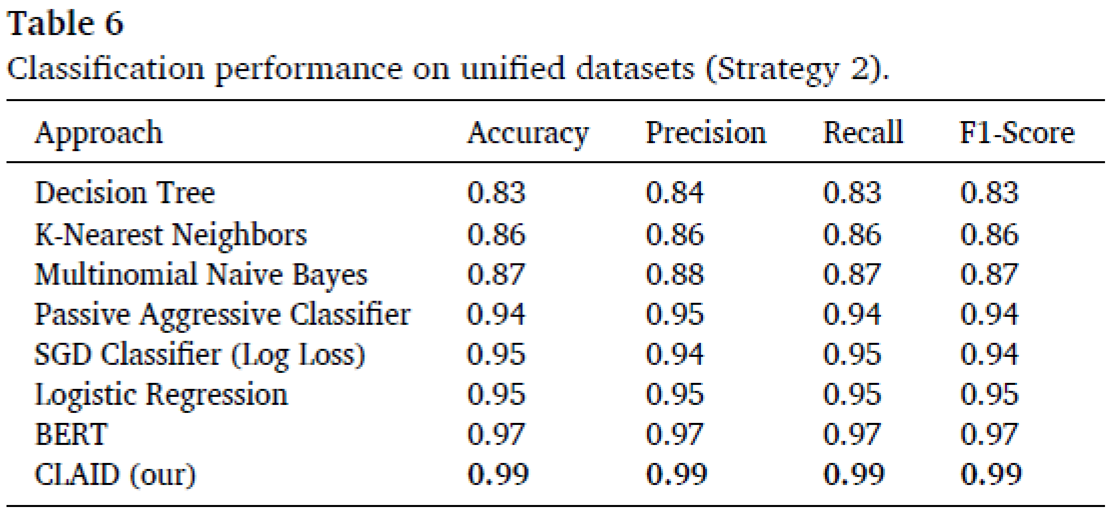


Table 6

Classification performance on unified datasets (Strategy 2).

Approach	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.83	0.84	0.83	0.83
K-Nearest Neighbors	0.86	0.86	0.86	0.86
Multinomial Naive Bayes	0.87	0.88	0.87	0.87
Passive Aggressive Classifier	0.94	0.95	0.94	0.94
SGD Classifier (Log Loss)	0.95	0.94	0.95	0.94
Logistic Regression	0.95	0.95	0.95	0.95
BERT	0.97	0.97	0.97	0.97
CLAID (our)	**0.99**	**0.99**	**0.99**	**0.99**

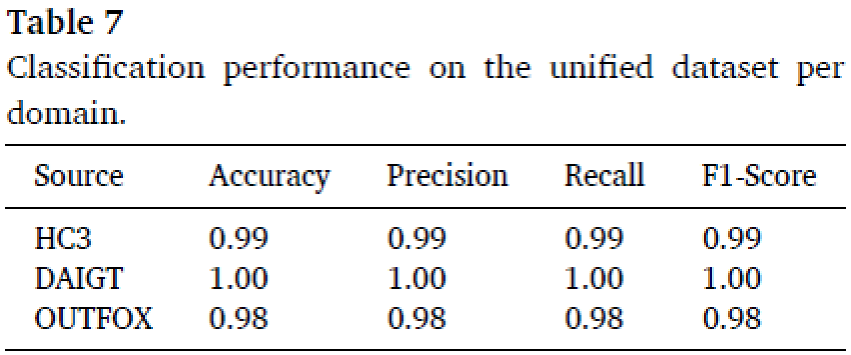


Table 7

Classification performance on the unified dataset per domain.

Source	Accuracy	Precision	Recall	F1-Score
HC3	0.99	0.99	0.99	0.99
DAIGT	1.00	1.00	1.00	1.00
OUTFOX	0.98	0.98	0.98	0.98

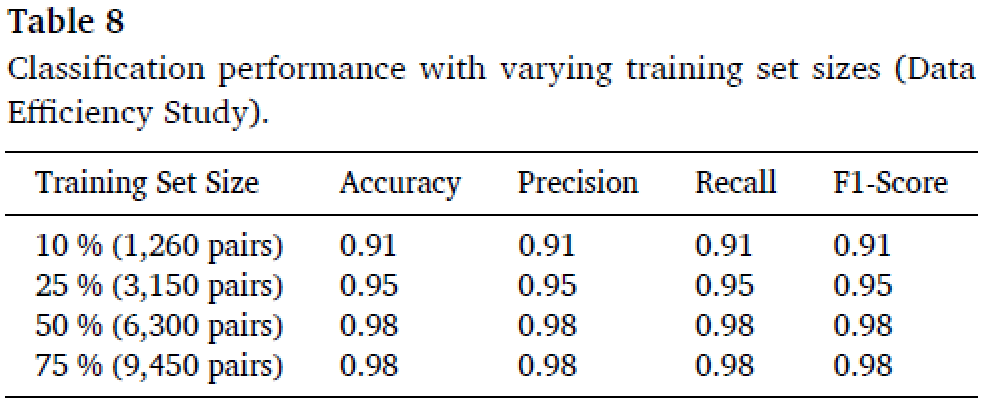


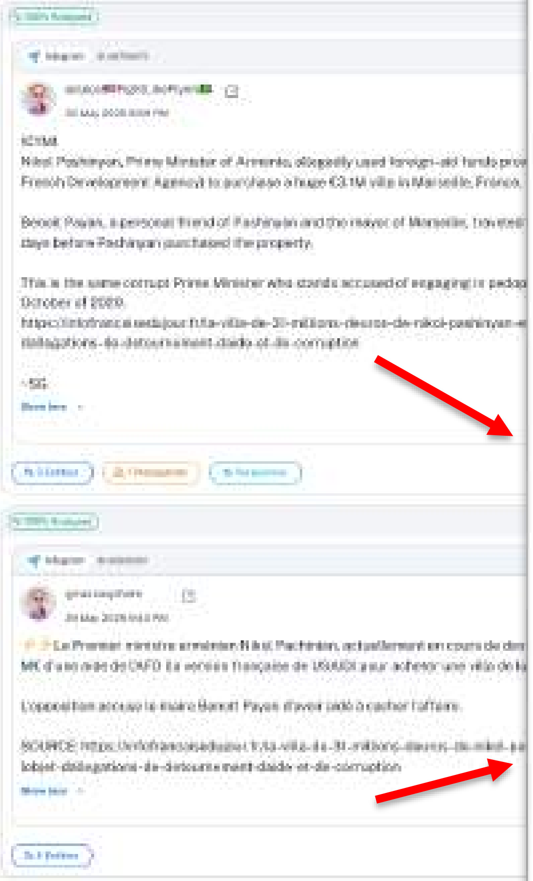
Table 8

Classification performance with varying training set sizes (Data Efficiency Study).

Training Set Size	Accuracy	Precision	Recall	F1-Score
10 % (1,260 pairs)	0.91	0.91	0.91	0.91
25 % (3,150 pairs)	0.95	0.95	0.95	0.95
50 % (6,300 pairs)	0.98	0.98	0.98	0.98
75 % (9,450 pairs)	0.98	0.98	0.98	0.98

Di Gisi, M., Fenza, G., Gallo, M., & Loia, V. (2025). Contrastive siamese network for detecting AI-generated text across domains and models. Neurocomputing, 131983.

AI-Generated Text Detection



ICYMI
Nikol Pashinyan, Prime Minister of Armenia, allegedly used foreign-aid funds pro
French Development Agency) to purchase a huge €3.1M villa in Marseille, France.

Benoit Payan, a personal friend of Pashinyan and the mayor of Marseille, traveled
days before Pashinyan purchased the property.

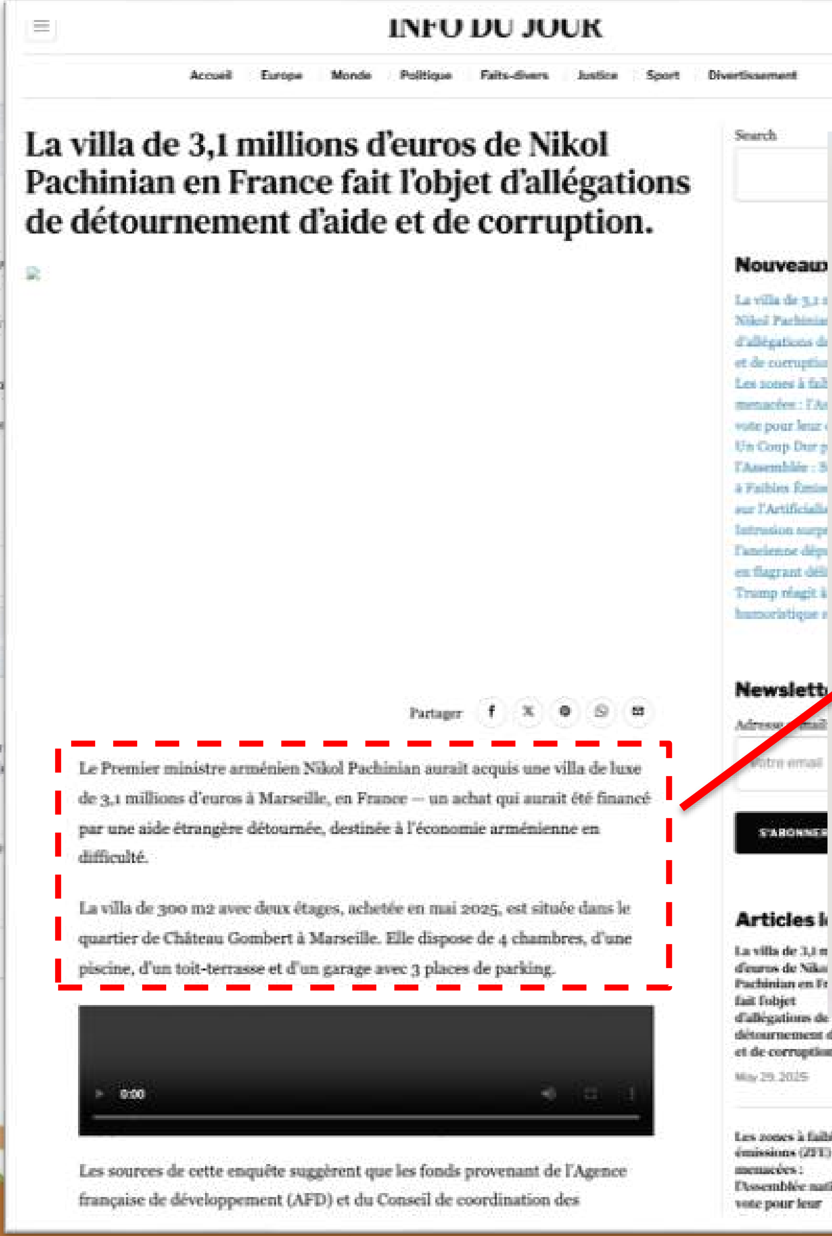
This is the same corrupt Prime Minister who stands accused of engaging in pedop
October of 2020.
https://infofrancaisedujour.fr/la-villa-de-31-millions-deuros-de-nikol-pashinyan-e
lallegations-de-detournement-daide-et-de-corruption

- SG

Le Premier ministre arménien Nikol Pachinian, actuellement en cours de dé
M€ d'une aide de l'AFD (la version française de USAID) pour acheter une villa de lu

L'opposition accuse le maire Benoît Payan d'avoir aidé à cacher l'affaire.

SOURCE: https://infofrancaisedujour.fr/la-villa-de-31-millions-deuros-de-nikol-pa
lobjet-dallegations-de-detournement-daide-et-de-corruption



La villa de 3,1 millions d'euros de Nikol Pachinian en France fait l'objet d'allégations de détournement d'aide et de corruption.

Le Premier ministre arménien Nikol Pachinian aurait acquis une villa de luxe de 3,1 millions d'euros à Marseille, en France — un achat qui aurait été financé par une aide étrangère détournée, destinée à l'économie arménienne en difficulté.

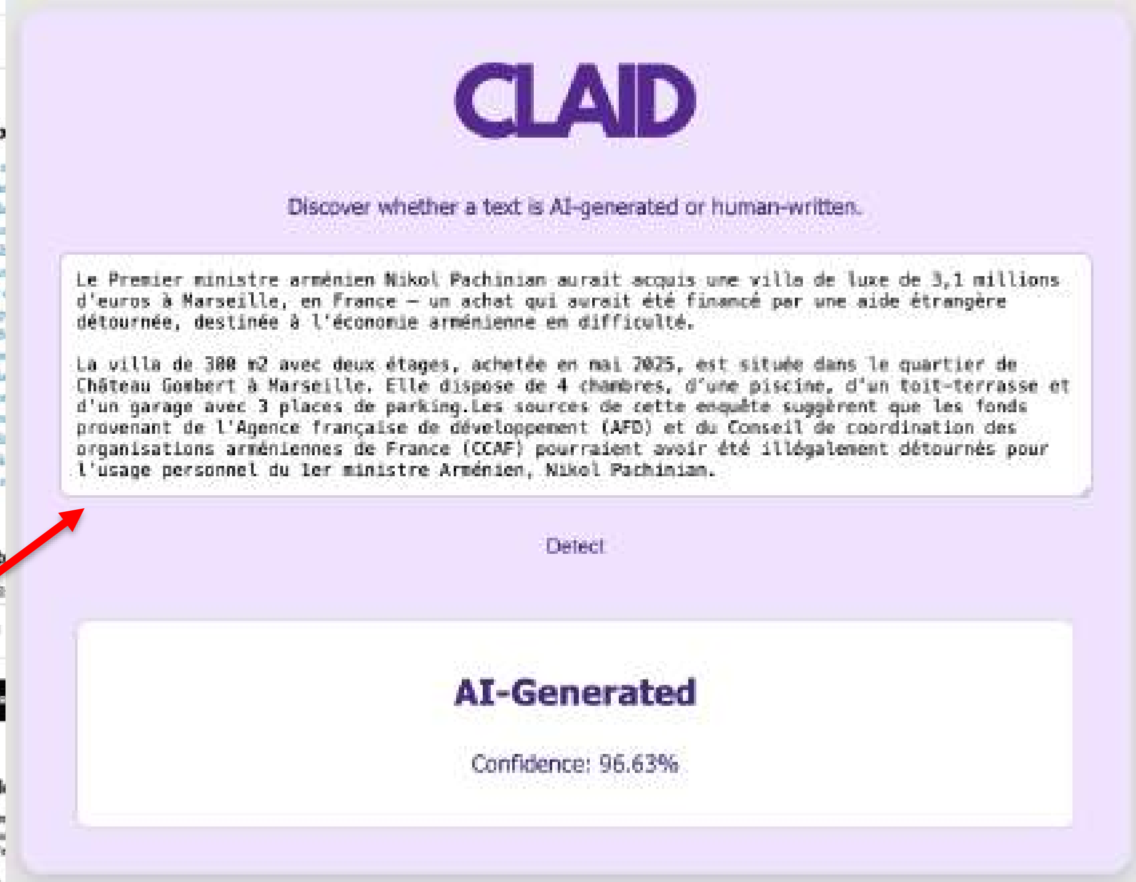
La villa de 300 m2 avec deux étages, achetée en mai 2025, est située dans le quartier de Château Gombert à Marseille. Elle dispose de 4 chambres, d'une piscine, d'un toit-terrasse et d'un garage avec 3 places de parking.

Les sources de cette enquête suggèrent que les fonds provenant de l'Agence française de développement (AFD) et du Conseil de coordination des

Nouveaux
La villa de 3,1 m
Nikol Pachinian
d'allégations de
et de corruption
Les zones à faib
menacées : l'As
vote pour leur
Un Coup Dur p
l'Assemblée : S
à Faibles Émiss
sur l'Artificialis
Intrusion surp
l'ancienne dépu
en flagrant dél
Trump réagit à
humoristique

Articles l
La villa de 3,1 m d'euros de Nikol Pachinian en Fr fait l'objet d'allégations de détournement d'aide et de corruption.
May 29, 2025

Les zones à faibles émissions (ZFE) menacées : l'Assemblée nationale vote pour leur



Le Premier ministre arménien Nikol Pachinian aurait acquis une villa de luxe de 3,1 millions d'euros à Marseille, en France – un achat qui aurait été financé par une aide étrangère détournée, destinée à l'économie arménienne en difficulté.

La villa de 300 m2 avec deux étages, achetée en mai 2025, est située dans le quartier de Château Gombert à Marseille. Elle dispose de 4 chambres, d'une piscine, d'un toit-terrasse et d'un garage avec 3 places de parking.Les sources de cette enquête suggèrent que les fonds provenant de l'Agence française de développement (AFD) et du Conseil de coordination des organisations arméniennes de France (CCAF) pourraient avoir été illégalement détournés pour l'usage personnel du 1er ministre Arménien, Nikol Pachinian.

AI-Generated

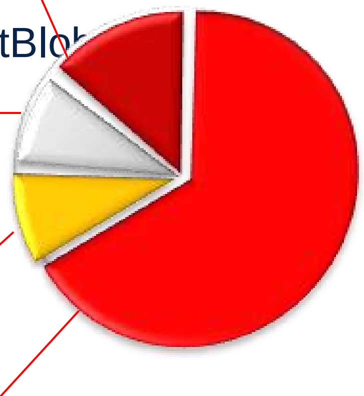
Confidence: 96.63%

Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What's next?
- **References –** Key sources and suggested readings

Credibility Scoring of News Outlets

- TEXT SCORE **20%**
  - Readability Score: The Flesch Reading Ease test evaluates the readability of the text.
  - Grammar Score: The grammatical structure of all web page content is computed by analyzing sentence structures.
  - Typo Score: Similarity is computed between the input text and its corrected version generated by TextBlo
- AMOUNT OF BANNERS **10%**
  - A multimodal LLaVA model (liuhaotian/llava-v1.5-7b) was used.
- TRAFFIC SCORE **10%**
  - Open PageRank API.
- CONTENT ANALYSIS THROUGH AI **60%**
  - Clickbait Headline Detection – christinacdl/XLM-RoBERTa-Clickbait-Detection-new – Accuracy 98%
  - Propaganda Detection – cstnz/PropagandaDetection – Accuracy 90%
  - Political Bias Detection – bucketre-search/politicalBiasBERT – Accuracy 72%
  - Fake News Evaluation – amzab/roberta-fake-news-classification – Accuracy 99%
- AUTHOR SCORE: not available yet



Experimentation Results – Newsguard Correlation

FakeNewsCorpus dataset

- 81 domains
- 50 web pages for each domain
- 5 weeks for each domain, on average

Spearman correlation: 81%

Pearson correlation: 84%.

MSCS has the following interpretation:

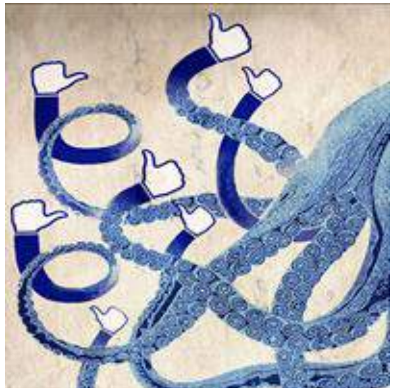
- 0 < MSCS < 39, the source is considered less credible, so it is necessary to proceed with extreme caution;
- 40 < MSCS < 59, the source is considered less credible, so it is necessary to proceed with caution;
- 60 < MSCS < 74, the source is credible, but with some exceptions;
- 75 < MSCS < 99, the source is generally reliable;
- MSCS = 100, the source is highly reliable.

Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What's next?
- **References –** Key sources and suggested readings

Countering Online Radicalization

Detecting & Reducing Online Radicalization



Berjawi, Omran, et al. "Mitigating radicalization in recommender systems by rewiring graph with deep reinforcement learning." Online Social Networks and Media 48 (2025): 100325.

Key takeaway: Radicalization can be measured, forecasted, and actively reduced through behavior-aware indicators and adaptive recommender interventions.

Role of Influential Actors in Opinion Dynamics

Analyzing the Persuasive Strategies of Influencers and News Media on Social Media. Omran Berjawi, Rida Khatoun and Giuseppe Fenza. To appear in the International Conference on Computer Systems and Applications (AICCSA 2025).

Key takeaway: Influencers shape polarization not only through network position, but through adaptive rhetoric aligned with audience behavior.

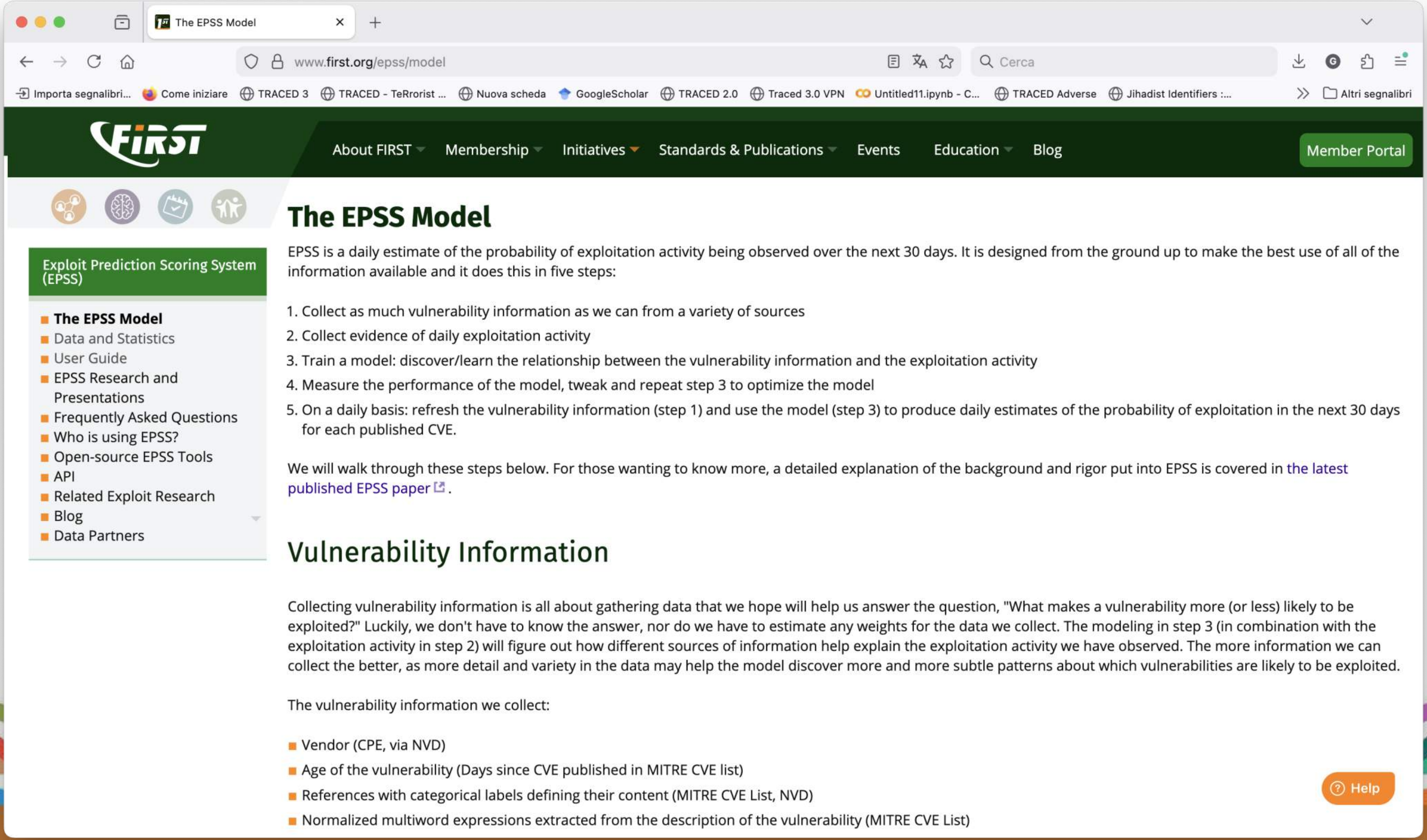
Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring, countering radicalization
- **Concluding Remarks –** What have we done? What's next?
- **References** – Key sources and suggested readings

From Debunking to Prebunking

- **Debunking**
- Debunking is a reactive strategy that aims to correct misinformation after it has already spread, by identifying false or misleading claims and replacing them with verified, accurate information.
- **Prebunking**
- Prebunking is a preventive strategy that aims to inoculate people against misinformation before they are exposed to it, by warning them about common manipulation techniques and misleading narratives.

Exploit Prediction Scoring System



The EPSS Model
www.first.org/epss/model
Cerca
Importa segnalibri... Come iniziare TRACED 3 TRACED - TeRrorist ... Nuova scheda GoogleScholar TRACED 2.0 Traced 3.0 VPN Untitled11.ipynb - C... TRACED Adverse Jihadist Identifiers :... Altri segnalibri
FIRST
About FIRST Membership Initiatives Standards & Publications Events Education Blog
Member Portal
Exploit Prediction Scoring System (EPSS)
The EPSS Model
Data and Statistics
User Guide
EPSS Research and Presentations
Frequently Asked Questions
Who is using EPSS?
Open-source EPSS Tools
API
Related Exploit Research
Blog
Data Partners
Help

The EPSS Model

EPSS is a daily estimate of the probability of exploitation activity being observed over the next 30 days. It is designed from the ground up to make the best use of all of the information available and it does this in five steps:

1. Collect as much vulnerability information as we can from a variety of sources
2. Collect evidence of daily exploitation activity
3. Train a model: discover/learn the relationship between the vulnerability information and the exploitation activity
4. Measure the performance of the model, tweak and repeat step 3 to optimize the model
5. On a daily basis: refresh the vulnerability information (step 1) and use the model (step 3) to produce daily estimates of the probability of exploitation in the next 30 days for each published CVE.

We will walk through these steps below. For those wanting to know more, a detailed explanation of the background and rigor put into EPSS is covered in the latest published EPSS paper .

Vulnerability Information

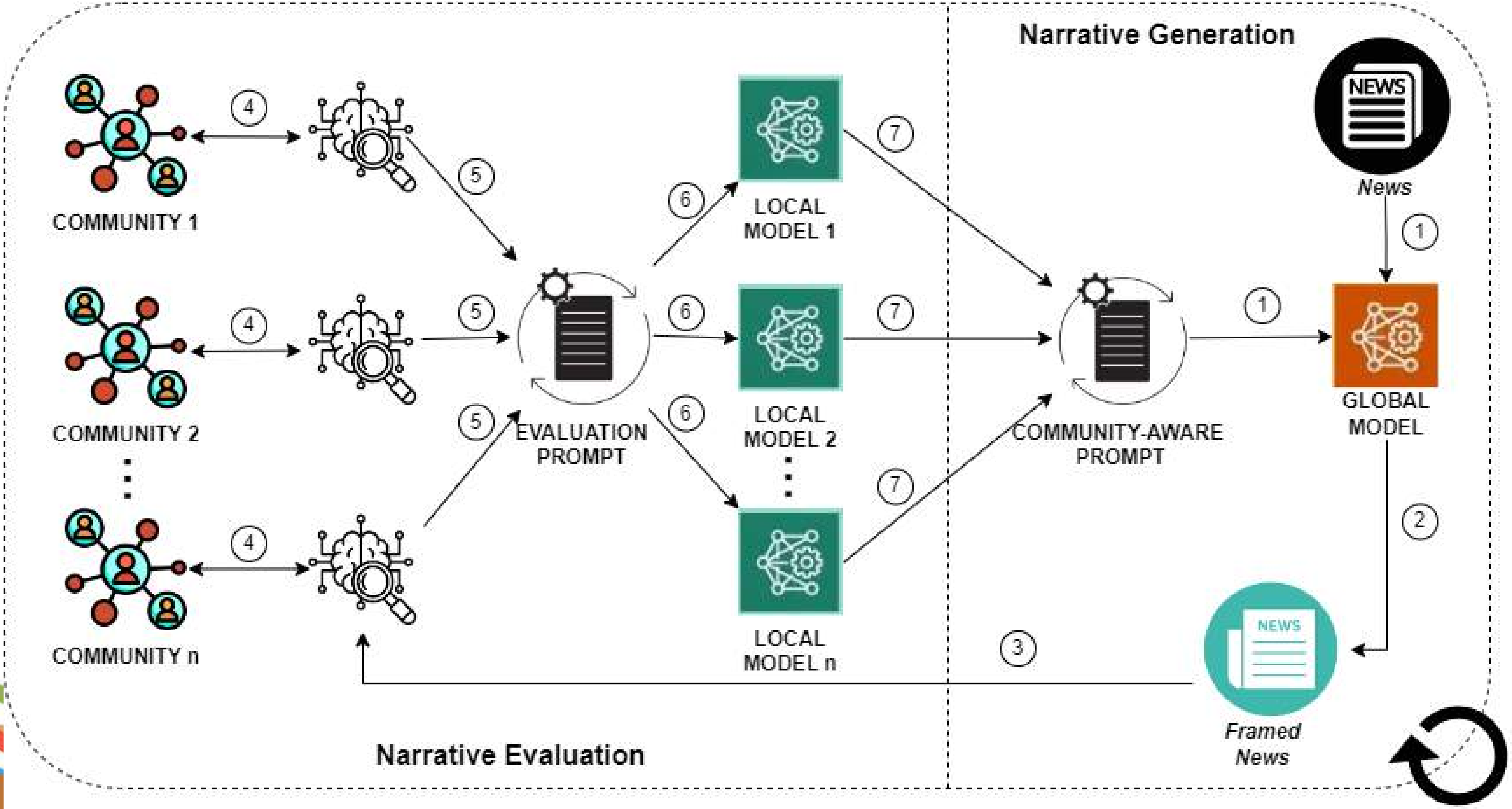
Collecting vulnerability information is all about gathering data that we hope will help us answer the question, "What makes a vulnerability more (or less) likely to be exploited?" Luckily, we don't have to know the answer, nor do we have to estimate any weights for the data we collect. The modeling in step 3 (in combination with the exploitation activity in step 2) will figure out how different sources of information help explain the exploitation activity we have observed. The more information we can collect the better, as more detail and variety in the data may help the model discover more and more subtle patterns about which vulnerabilities are likely to be exploited.

The vulnerability information we collect:

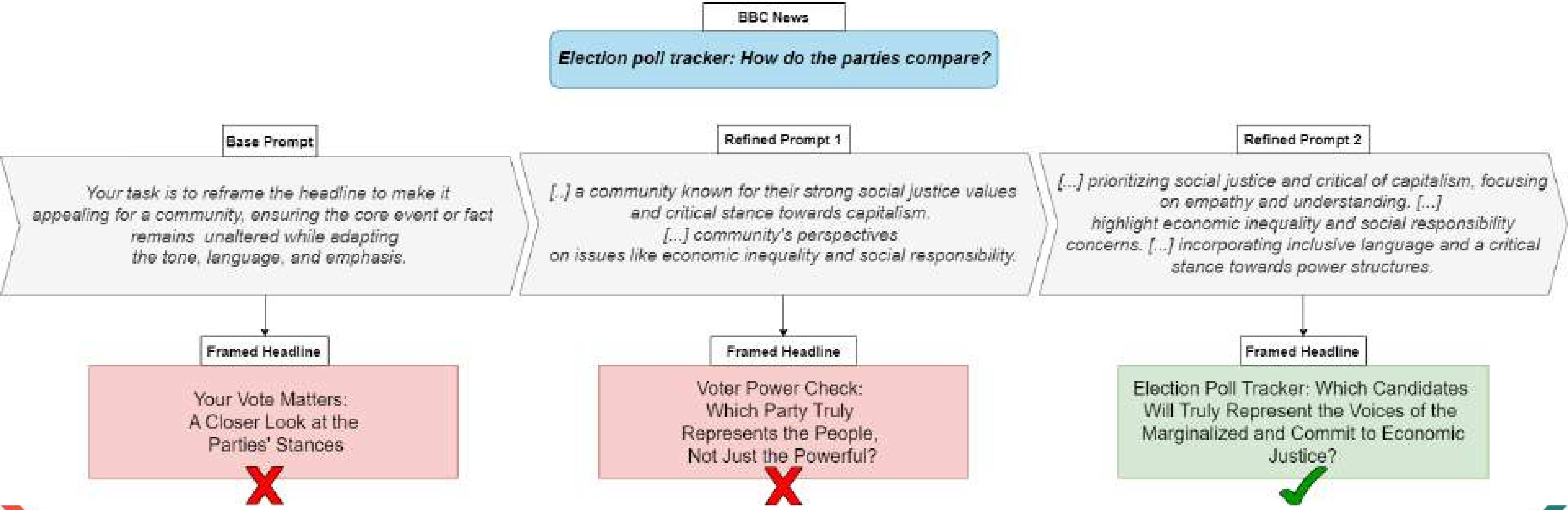
- Vendor (CPE, via NVD)
- Age of the vulnerability (Days since CVE published in MITRE CVE list)
- References with categorical labels defining their content (MITRE CVE List, NVD)
- Normalized multiword expressions extracted from the description of the vulnerability (MITRE CVE List)

Cognitive Vulnerability Exploitation Score

Framework Overview



Prompt Iterative Refinement



BBC News

Election poll tracker: How do the parties compare?

Base Prompt

Your task is to reframe the headline to make it appealing for a community, ensuring the core event or fact remains unaltered while adapting the tone, language, and emphasis.

Framed Headline

Your Vote Matters: A Closer Look at the Parties' Stances ✗

Refined Prompt 1

[..] a community known for their strong social justice values and critical stance towards capitalism. [...] community's perspectives on issues like economic inequality and social responsibility.

Framed Headline

Voter Power Check: Which Party Truly Represents the People, Not Just the Powerful? ✗

Refined Prompt 2

[...] prioritizing social justice and critical of capitalism, focusing on empathy and understanding. [...] highlight economic inequality and social responsibility concerns. [...] incorporating inclusive language and a critical stance towards power structures.

Framed Headline

Election Poll Tracker: Which Candidates Will Truly Represent the Voices of the Marginalized and Commit to Economic Justice? ✓

Vulnerability Assessment

🔴 Highly Weak	⚪ Weak	🟡 Barely Strong	🟢 Strong
$i = 1$	$1 < i \leq \lfloor t/2 \rfloor$	$\lfloor t/2 \rfloor < i \leq t$	$i > t$

From LLM Agents to Social Media Digital Twins

Goal: Understand how human-driven influence shapes collective opinion beyond algorithms.

1. What We Did (Recent Work: LLM Agents)

- Studied opinion dynamics in networks of LLM-driven agents.
- Showed bias amplification:
  - Even a small fraction of biased agents shifts collective opinion.
  - Leads to extremity convergence, not balanced consensus.
- Highlighted risks of deploying LLM agents in social simulations and decision-making.

2. What We Will Do (Future Research Agenda)

- Develop Social Media Digital Twins: Virtual replicas of real online platforms
- Key components:
  - Graph-based social networks
  - LLM-driven user agents
  - Platform-level behavioral and recommendation rules
- Enable: Safe testing of interventions (e.g., recommender rewiring, influencer moderation)
- Bridge computational social science, AI safety, and platform governance

Key takeaway: A unified experimental framework to study radicalization, influence, and AI-mediated opinion dynamics before real-world deployment.

Conclusions

- Shifting our focus to the prebunking area
- Focus on Technology Transfer:
  - Filing Patents
  - Starting Pilots with National Institutions
- Project Open Repositories
  - **Source Code (GitHub):**

    https://github.com/Information-Disorder-Awareness
  - **Models & Resources (Hugging Face):**

    https://huggingface.co/IDA-SERICS
- References:
  - https://scholar.google.com/citations?hl=it&user=0C3IjEIAAAAJ&view_op=list_works&sortby=pubdate

SERICS
SECURITY AND RIGHTS IN THE CYBERSPACE

HIPPOCRATICA CIVITAS
STUDIUM SALERNI

Merci!
Grazie!
Thank you!
Gracias!
Danke!

Agenda

- **Why It Matters** – The impact of disinformation on society
- **Countering Disinformation @ UNISA –** SIEM-like Platform
- **Research Activities**: benchmarking, fact-checking, generated content detection, credibility scoring of news outlets
- **Concluding Remarks –** What have we done? What’s next?
- **References –** Key sources and suggested readings

Bibliography

- *Di Gisi, M., Fenza, G., Gallo, M. and Loia, V., 2025. Contrastive siamese network for detecting AI-generated text across domains and models. Neurocomputing, p.131983.*
- *Berjawi, O., Fenza, G., Khatoun, R. and Loia, V., 2025. Mitigating radicalization in recommender systems by rewiring graph with deep reinforcement learning. Online Social Networks and Media, 48, p.100325.*
- *Fenza, G., Furno, D., Gallo, M., Loia, V. and Trotta, P.P., 2024, September. Claim Verification Leveraging In-context Learning and Retrieval Augmented Generation. In International Conference on Advances in Social Networks Analysis and Mining (pp. 17-31). Cham: Springer Nature Switzerland.*
- *Fenza, G., Loia, V., Stanzione, C. and Di Gisi, M., 2024. Robustness of models addressing Information Disorder: A comprehensive review and benchmarking study. Neurocomputing, 596, p.127951.*
- *Berjawi, O., Khatoun, R. and Fenza, G., 2025, May. Digital Persuasion: Understanding the Impact of Online Influencers on Public Opinion. In International Conference on Persuasive Technology (pp. 117-127). Cham: Springer Nature Switzerland.*
- *Fenza, G., Gallo, M., Loia, V. and Stanzione, C., 2024, May. Evaluating Web Domain Credibility: A Multifactorial Score for Analyzing Online Reliability. In 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS) (pp. 1-8). IEEE.*
- *Maria Di Gisi, Giuseppe Fenza, Domenico Furno, Mariacristina Gallo, Vincenzo Loia, Pio Pasquale Trotta: Federated Prompt Tuning for News Framing: A Community-Aware Approach to Narrative Exploitability. IJCNN 2025: 1-8*
- https://scholar.google.com/citations?hl=it&user=0C3IjEIAAAAJ&view_op=list_works&sortby=pubdate