trueflag

who do you trust?

Automated detection of disinformation campaigns targeting brands

“How to turn automated fact-checking into a profitable business”

TrueFlag.ai is Newtral's tech spin-off to leverage our work on automated fact-checking

Nevtral

TV content | Fact-checking | AI

- **2013**: First team to introduce fact-checking on a prime time TV show in Spain
- **2017**: ‘El Objetivo’ is the first Spanish signatory of the International Fact-Checking Network
- **2018**: Meta Partnership (FB, Instagram)
- **2020**: Whatsapp Partnership
- **2021**: TikTok Partnership
- **2021**: Google collaboration to fight disinformation against vaccines

FACT-CHECKING PROCESS

Automated by Newtral

1 MONITORING

What is being distributed?

Twitter, Facebook, TV, Radio, Digital news...

2 SPOT FACTS

Is this verificable?
Is it **relevant**? **Dangerous**? Is it a joke?
Does it make sense?

3 VERIFICATION

Are there **trustable** sources?
Is it **fake**? Half true?
Can we verify quicker?

4 EXPLOITATION

How can we reach the same **audience** impacted by the fake?

TECH EVOLUTION

- **2018:** Linguistic rules and NLP frameworks
- **2019**: Traditional ML
- **2020:** DL monolingual (first BERTs)
- **2021:** DL multilingual (novel BERTs)
- **2022:** Sentence Transformers | T5 models
- **2023:** GenAI & vector DB
- **2024**: Agents
- **2025**: Multimodal

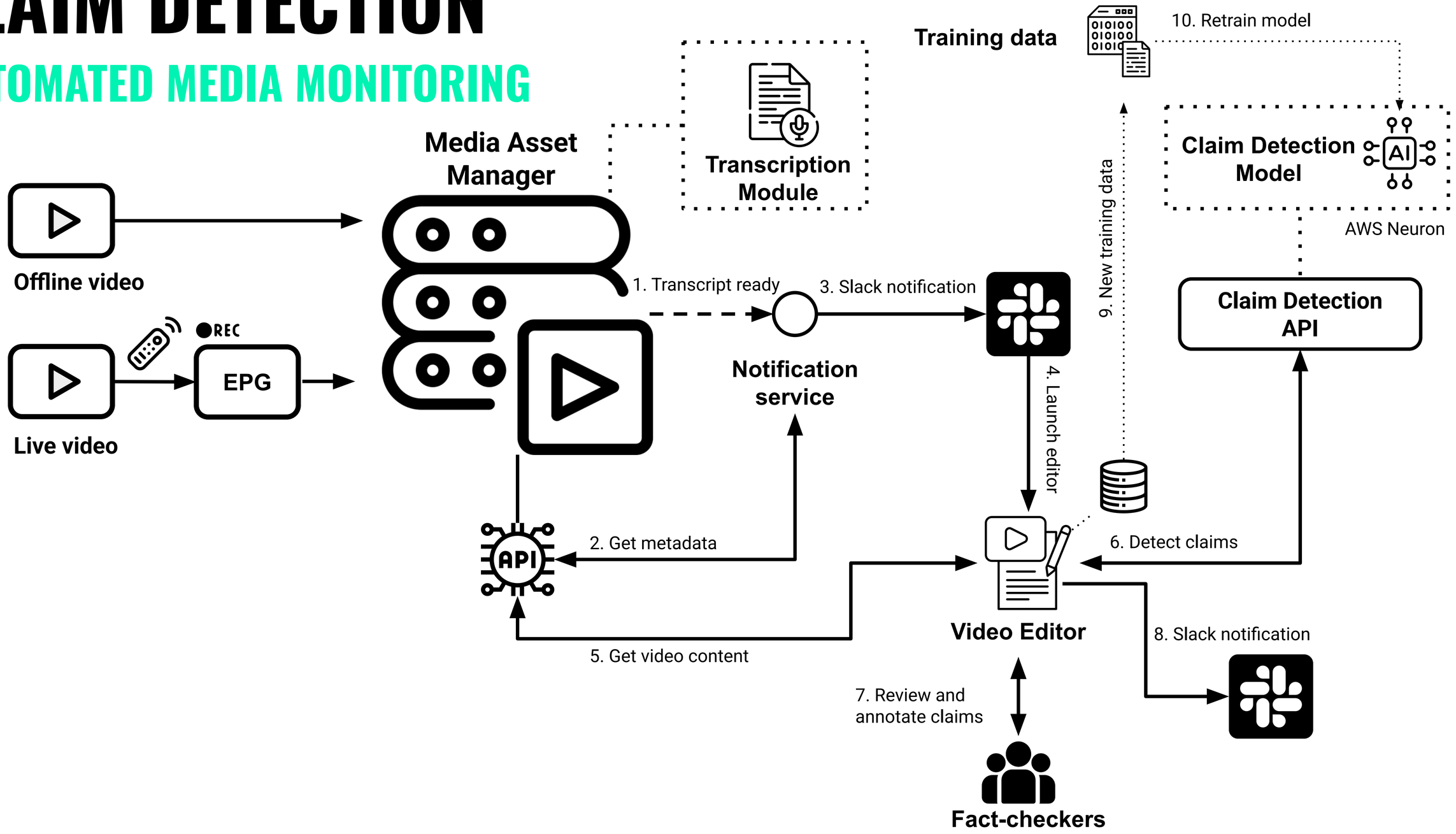
trueflag
who do you trust?

TECH STACK

	PYTHON	TYPESCRIPT
TEST	PyTest	Jest Mocha Supertest Chai
API REST	FastAPI Django	NestJS NodeJS React NextJS
AI	PyTorch Pandas BERT models SLMs / LLMs MLFLow SageMaker LangGraph Langfuse	
DATA	PowerBI Azure Data Lake Airflow AWS Glue Azure Data Factory Metabase	Playwright Pupeteer Kafka Events Socket.io
ORMs DB	PyMongo Alembic SQLAlchemy PySpark TypeORM Prisma MongoDB MariaDB Postgres Clickhouse Redis ElasticSearch Qdrant	
DEV TOOLS	TorchServe Triton Server Cloudformation Github Actions ECS Poetry Virtualenv Npm Terraform AWS Neuron Bitbucket pipelines Docker ELB Pyenv Yarn Nvm	

CLAIM DETECTION

AUTOMATED MEDIA MONITORING



CLAIM DETECTION

VIDEO EDITOR

60 min video

|

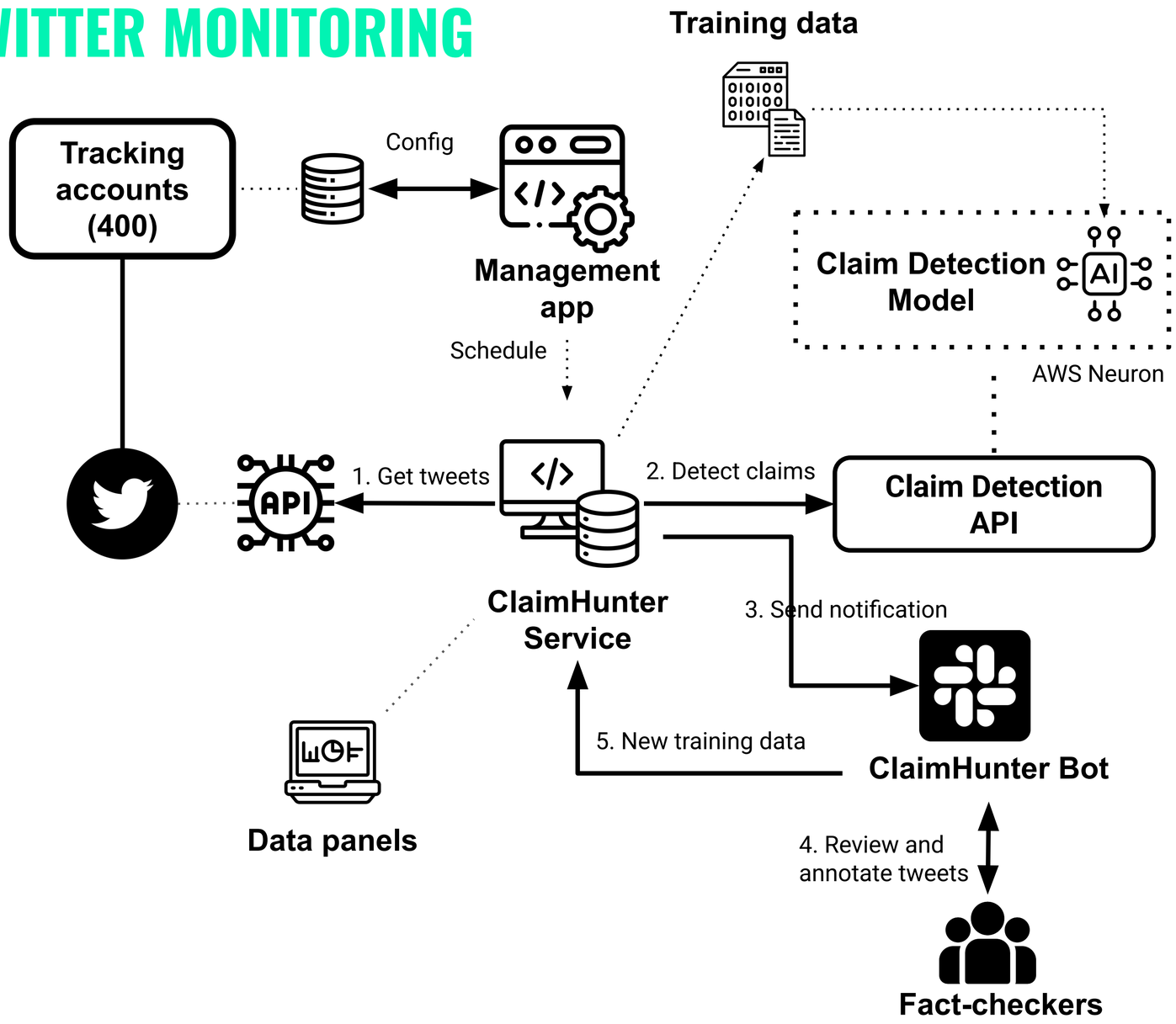
3 sec detection

|

5 min review

CLAIM DETECTION

TWITTER MONITORING

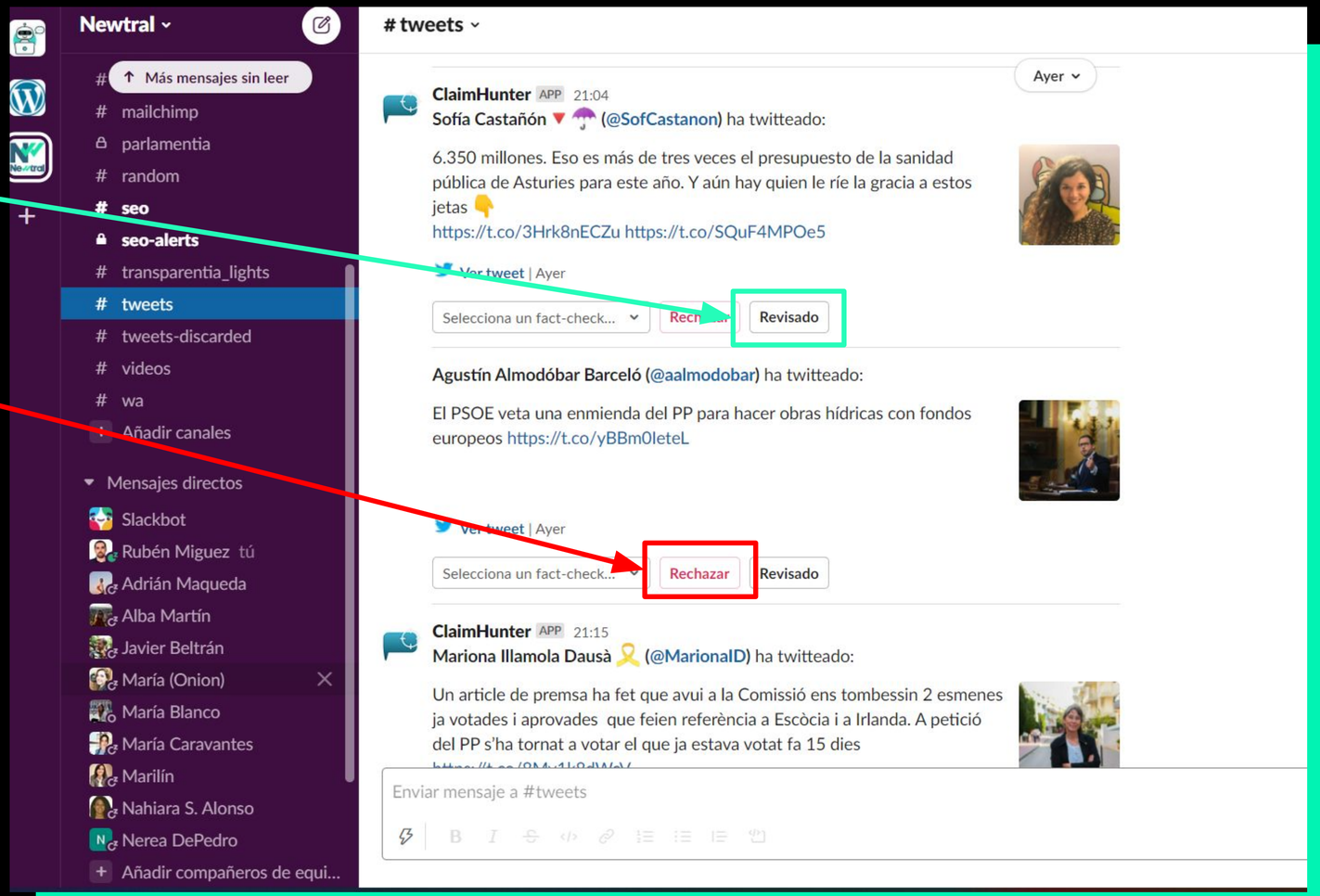


CLAIM DETECTION

TWITTER MONITORING

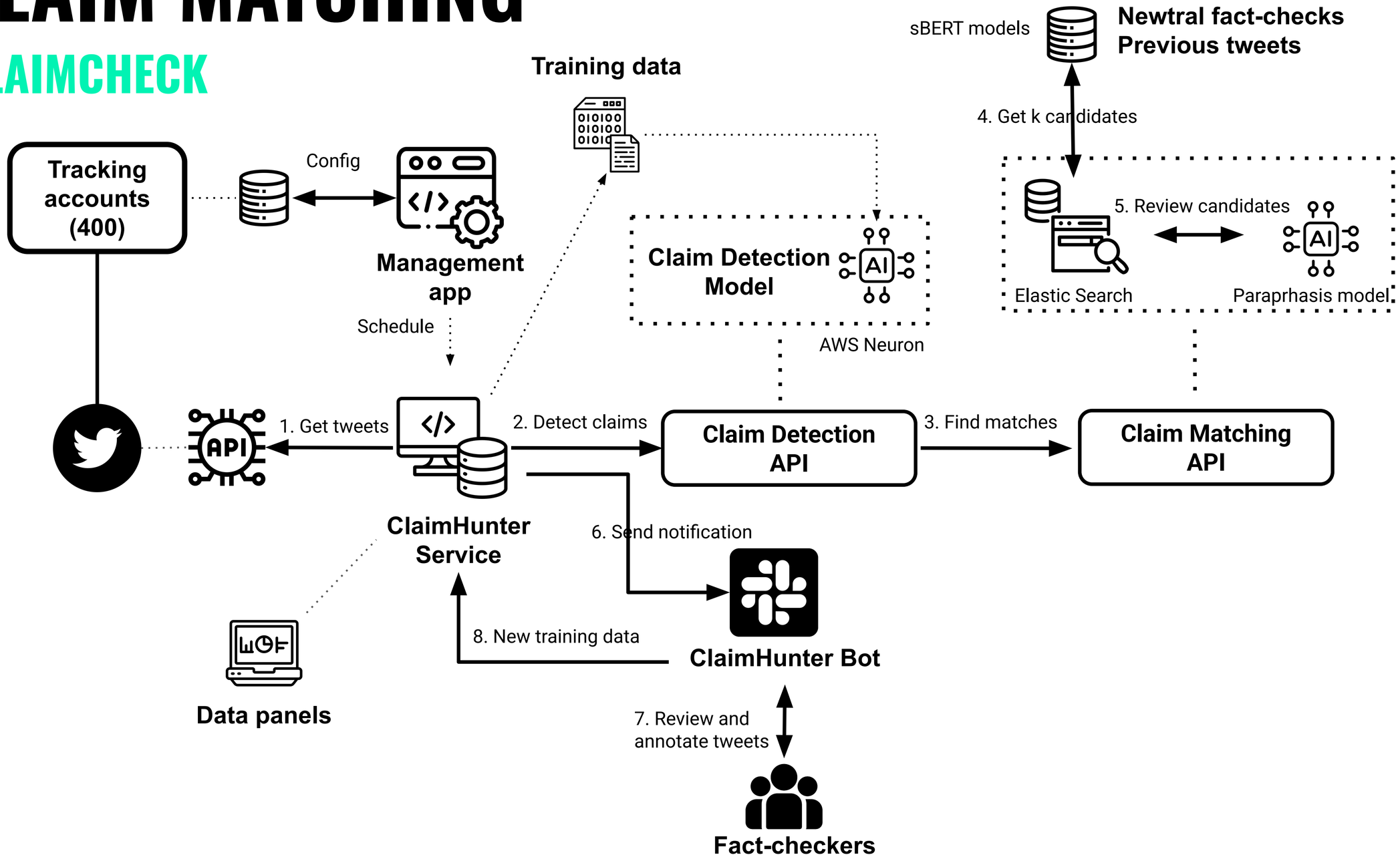
Positive feedback

Negative feedback



CLAIM MATCHING

CLAIMCHECK



Focusing on the user: a Chrome extension for detecting unreliable content

DISINFO PATTERNs

TRUST SCORE

4 BERT models:

- Toxicity
- Subjetivity
- Factuality
- Clickbait

TrustScore for evaluating content reliability:

- A - regular content
- E - low trust

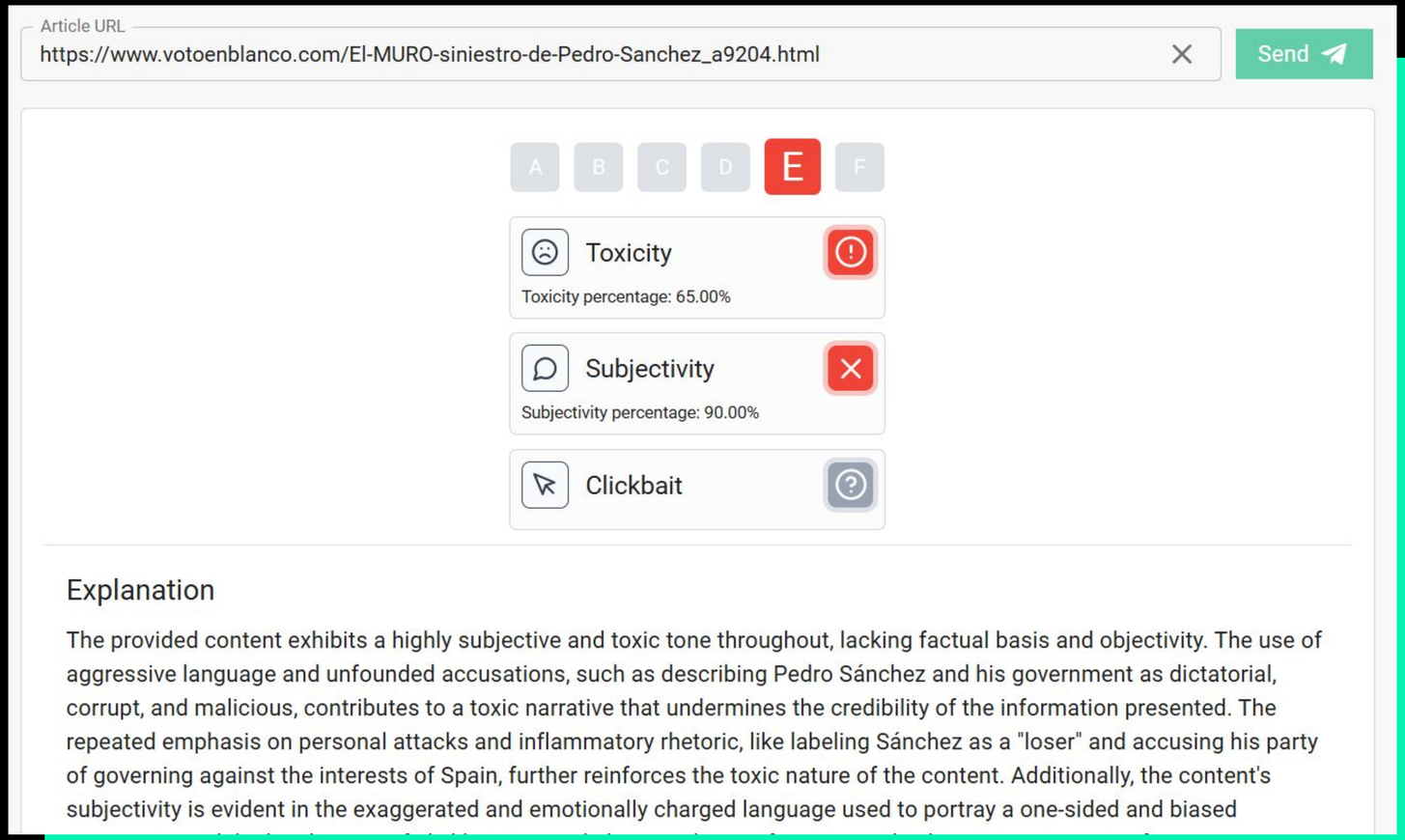
DISINFO PATTERNs

TRUST SCORE - GENAI

Trust Score explainability

LLMs for weak labeling:

- Saving annotation resources
- Quick model iteration
- Batch processing



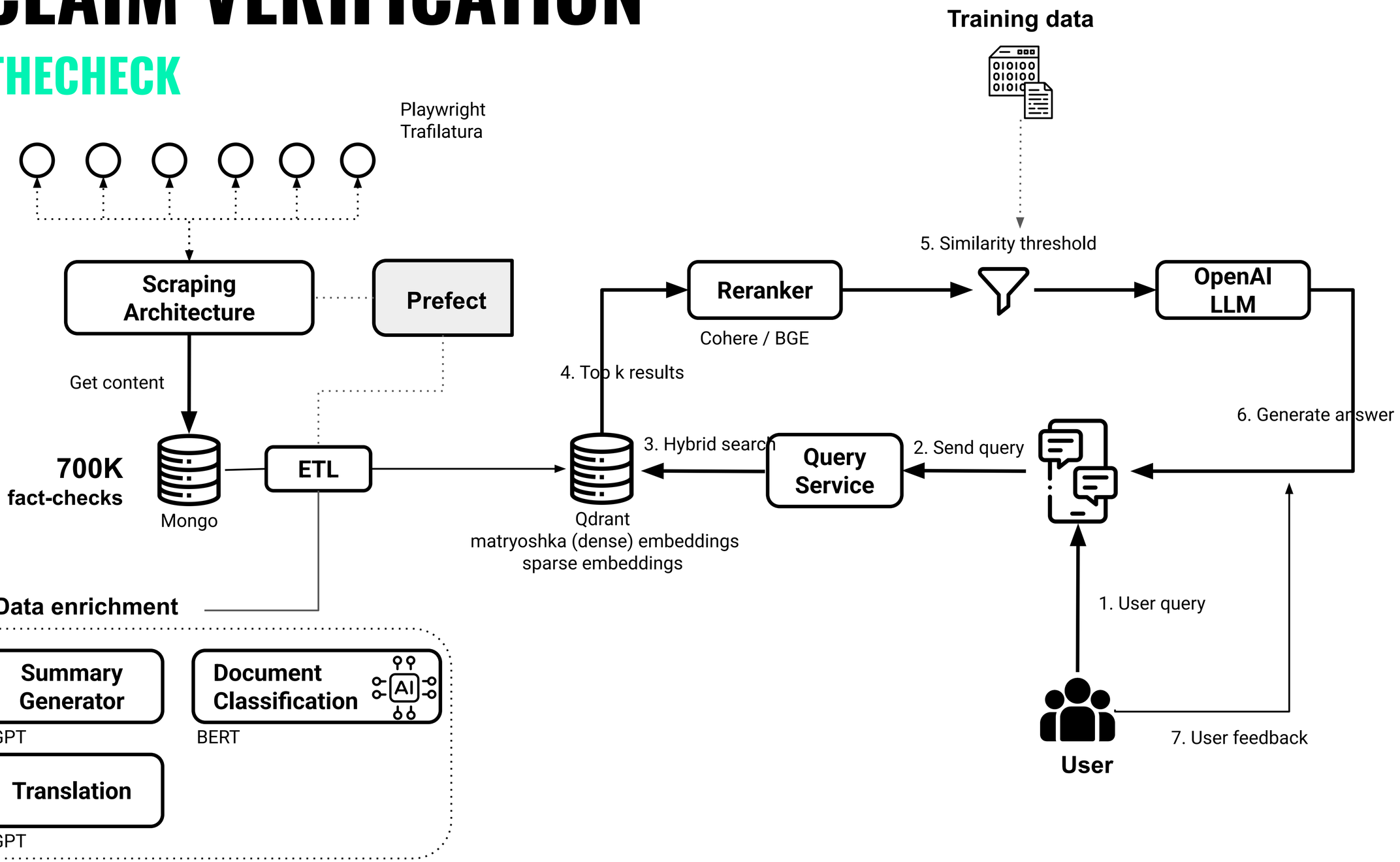
ruben.miguez@newtral.es

agustin.canas@newtral.es

Focusing on the user: a RAG chatbot for verifying false statements

CLAIM VERIFICATION

THECHECK



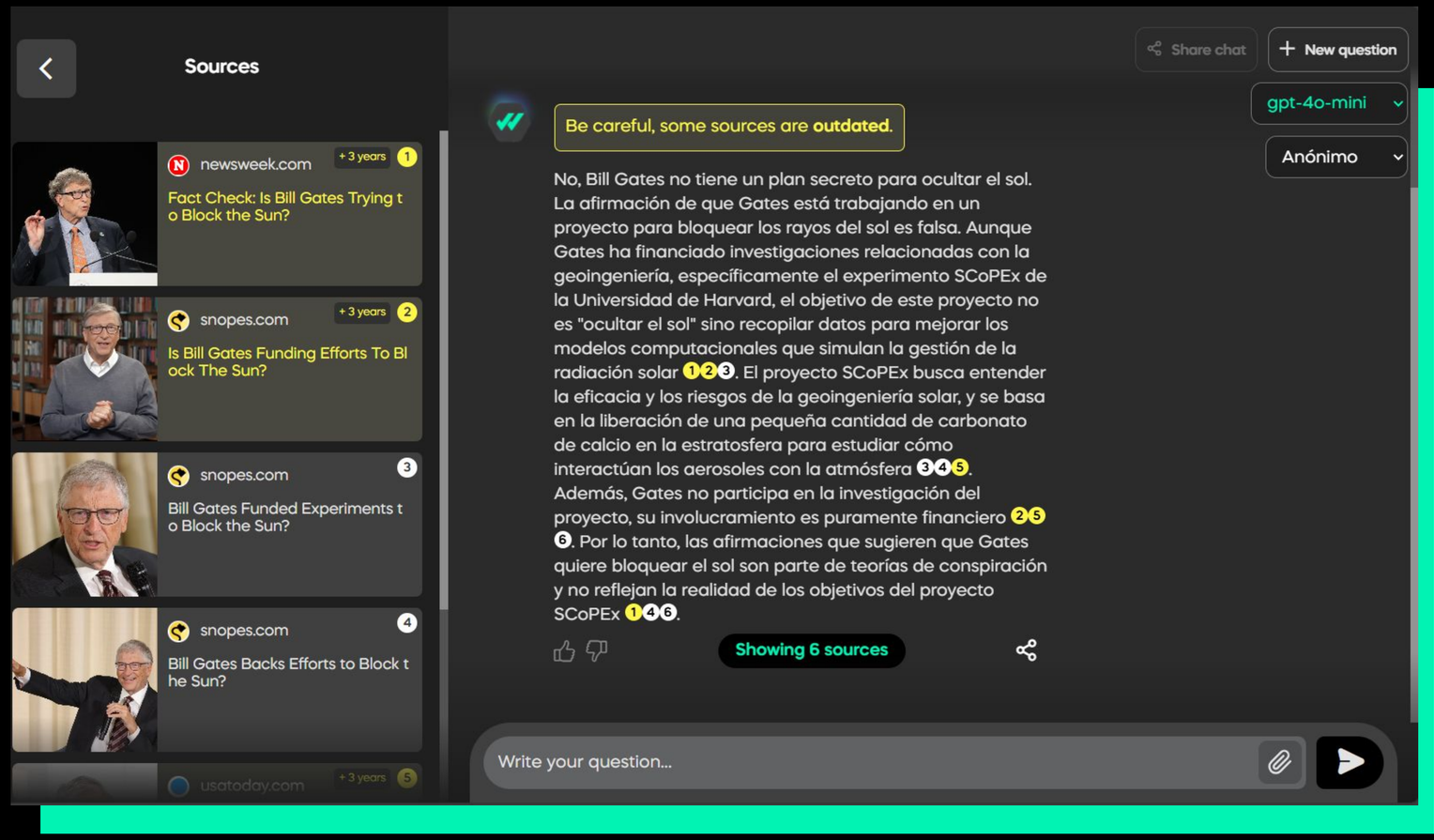
CLAIM VERIFICATION

THECHECK

Multilingual chatbot

Trusted sources only

Summarizes key points in a clear and easy to follow language



Detection of disinformation campaigns for brand protection

OUR SOLUTION

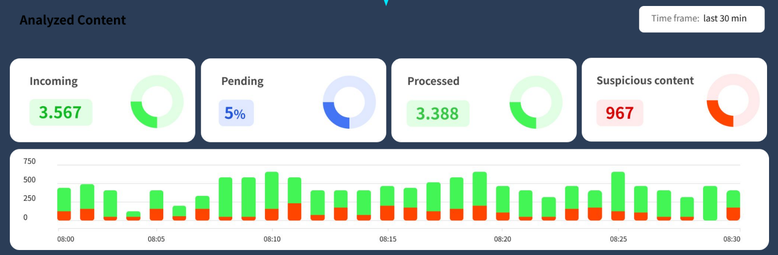
trueflag

who do you trust?

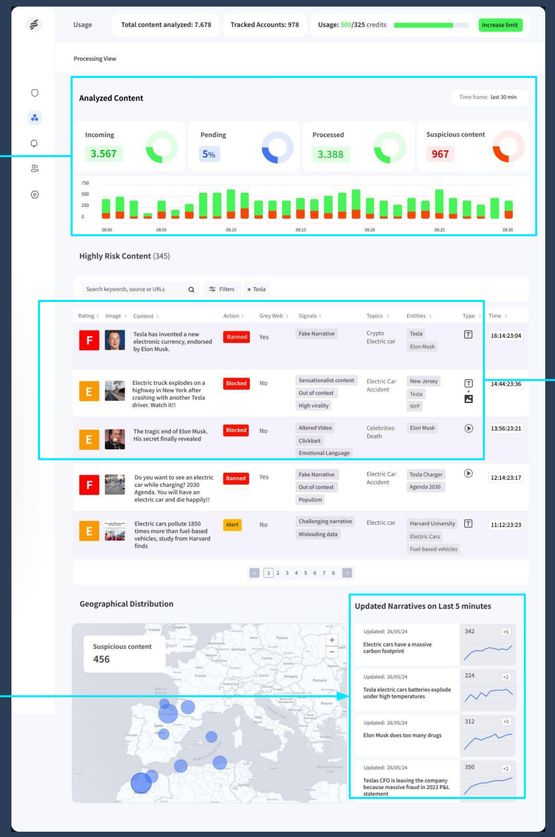
The 1st multimodal multilingual SaaS designed to automate the detection verification of disinformation in online content

UPSCALES TRUST

Verification process and alerts



Narratives



Early alerts
Predictive behavior

Ratings

Grey Web alerts and disinformation signals

Rating	Image	Content	Action	Grey Web	Signals	Topics	Entities	Type
F		Tesla has invented a new electronic currency, endorsed by Elon Musk.	Banned	Yes	Fake Narrative	Crypto Electric car	Tesla, Elon Musk	T
E		Electric truck explodes on a highway in New York after crashing with another Tesla driver. Watch it!!	Blocked	No	Sensationalist content, Out of context, High virality	Electric Car Accident	New Jersey, Tesla, NYP	T +
E		The tragic end of Elon Musk. His secret finally revealed	Blocked	No	Altered Video, Clickbait, Emotional Language	Celebrities Death	Elon Musk	

DECISION

THE PROBLEM

A threat increased by generative AI

More ability to create new content
Easier to disseminate content
No anti-disinfo policies on social networks

Interference through AI deep fakes and physical intimidation in India's elections 2024.

Instagram image of a nude woman bearing semblance to a public figure from India. Other, in a Facebook group, resembling an American figure in a compromising pose.

Hong Kong multinational firm loses HK$200 million after scammers stage deepfake video meeting.

WHY NOW?

Regulation on disinformation is being imposed[1]

Misinformation/disinformation is estimated as the top global risk over the short term for the second year in a row[2]

[1] EU's Digital Service Act and Code of Practice on Disinformation
[2] Global Risks Report 2025. WEF

THE PROBLEM

90%

of online content will be **generated** or assisted by AI tools by 2026

60%

of LLM's **fail** to provide accurate URLs and proper citations in 2025. Even if it % improves, they aren't a reliable source.

70%

More possibilities that people **shares** fake news that true news

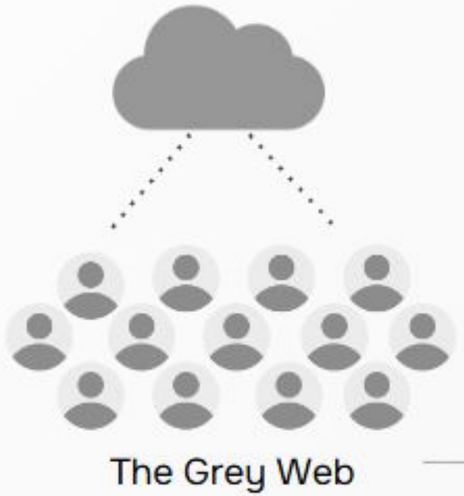
1st

concern among the business community. Disinformation is the global risks over extreme weather and state based armed conflict

trueflag

GENERAL OVERVIEW

1. TARGETED MONITORING



2. DETECTION | 3. VERIFICATION

trueflag

Verification SaaS
by Inaria

CLAIM DETECTION

CLAIM MATCHING

CLAIM VERIFICATION

DISINFO PATTERNS

Fake News

Ground Truth

Verified content + Client data + Other sources

VERIFICATION SAAS API

4. EXPLOITATION

trueflag for brands

Brand protection

- detection of disinformation narratives
- predictive alerts
- AI generated content
- advisory reports

Content moderation

- rating for regulated platforms
- ads analysis
- content safety
- chrome extension
- clickbait-free contents
- conversational chat

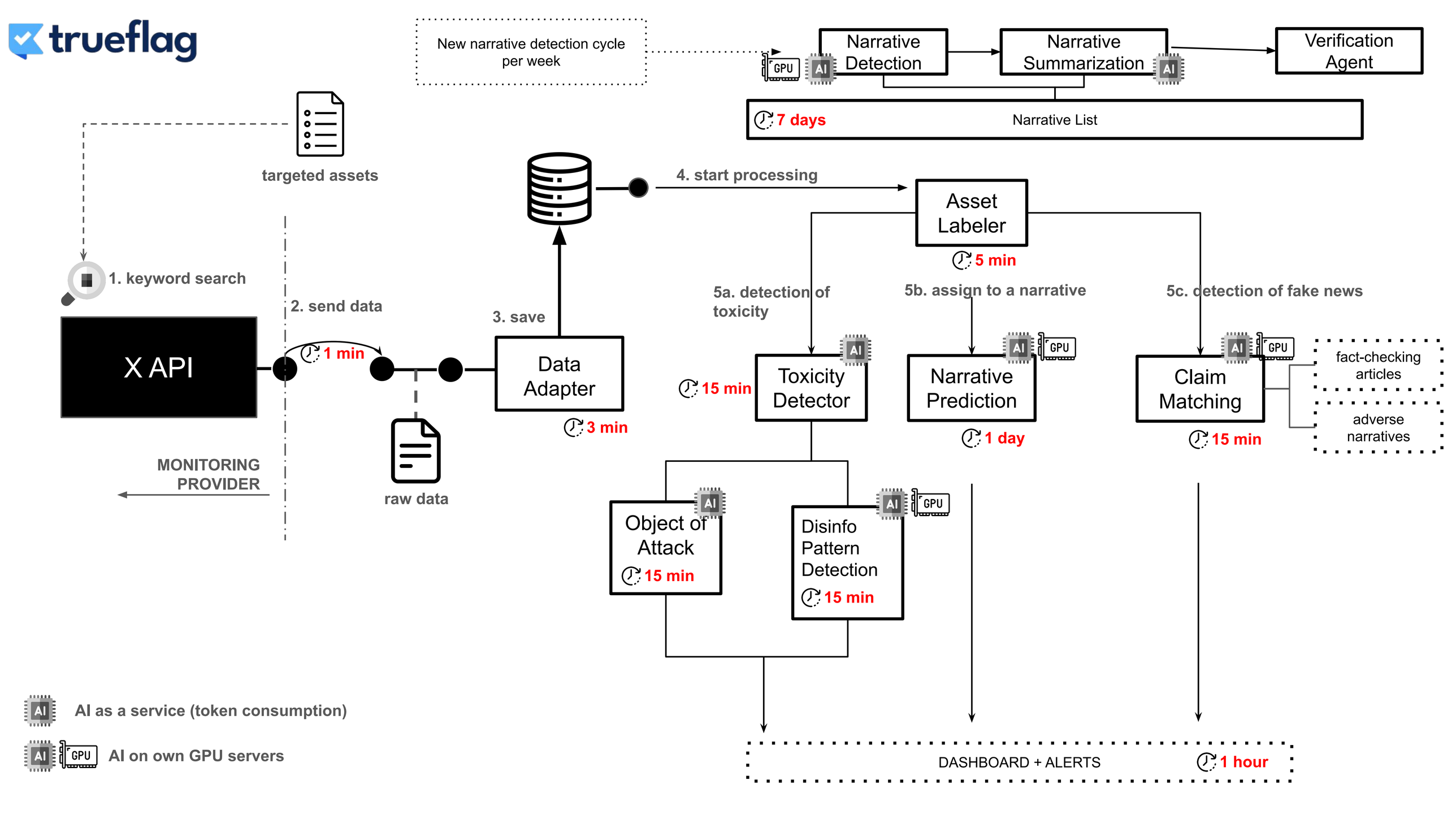
Phase 1: Set-up

Evaluate disinformation risk

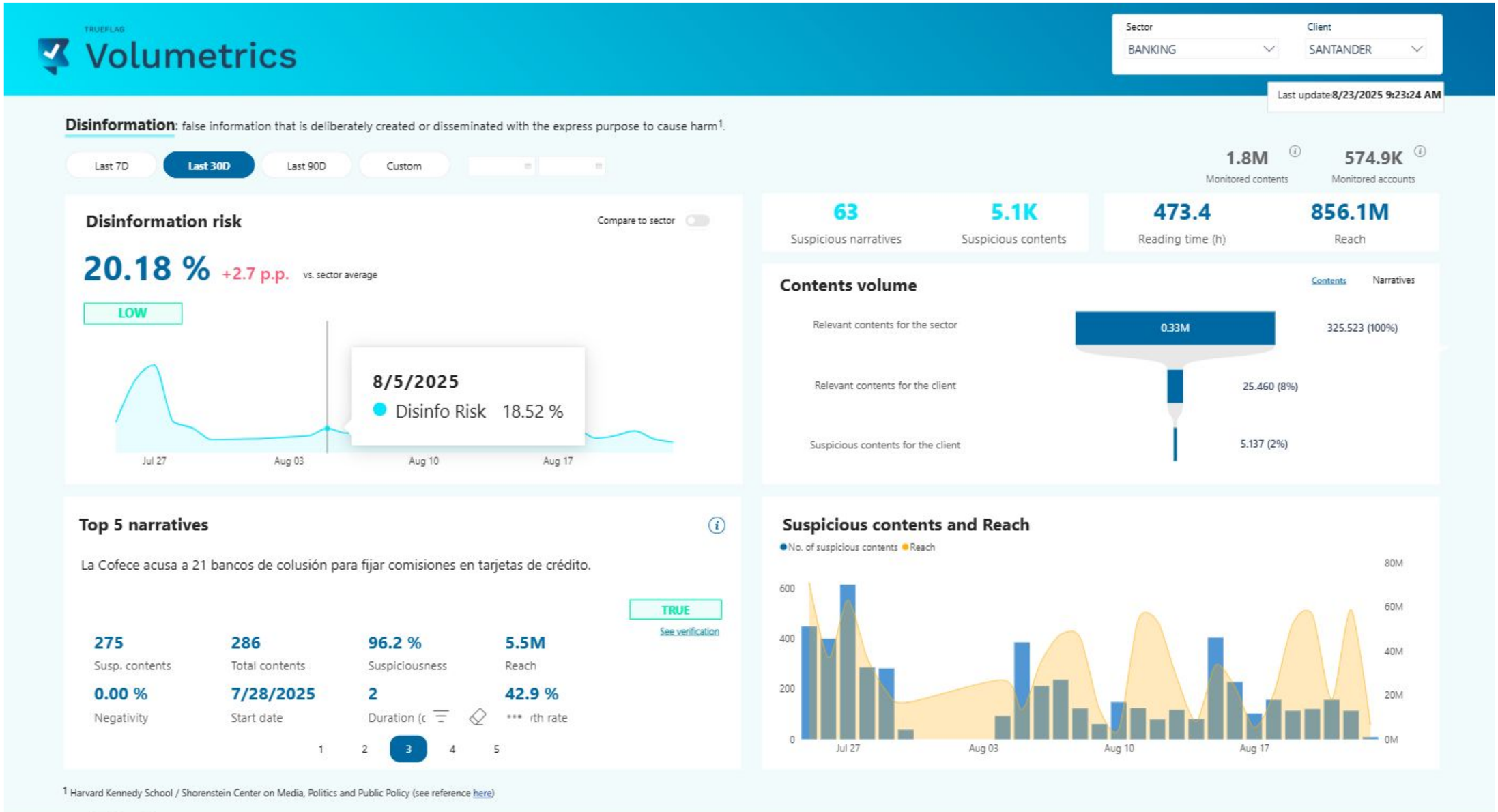
- **Assets** to be protected
- Definition of **Grey Web**
- Measure disinformation reach

Evaluate automation of disinformation analysis

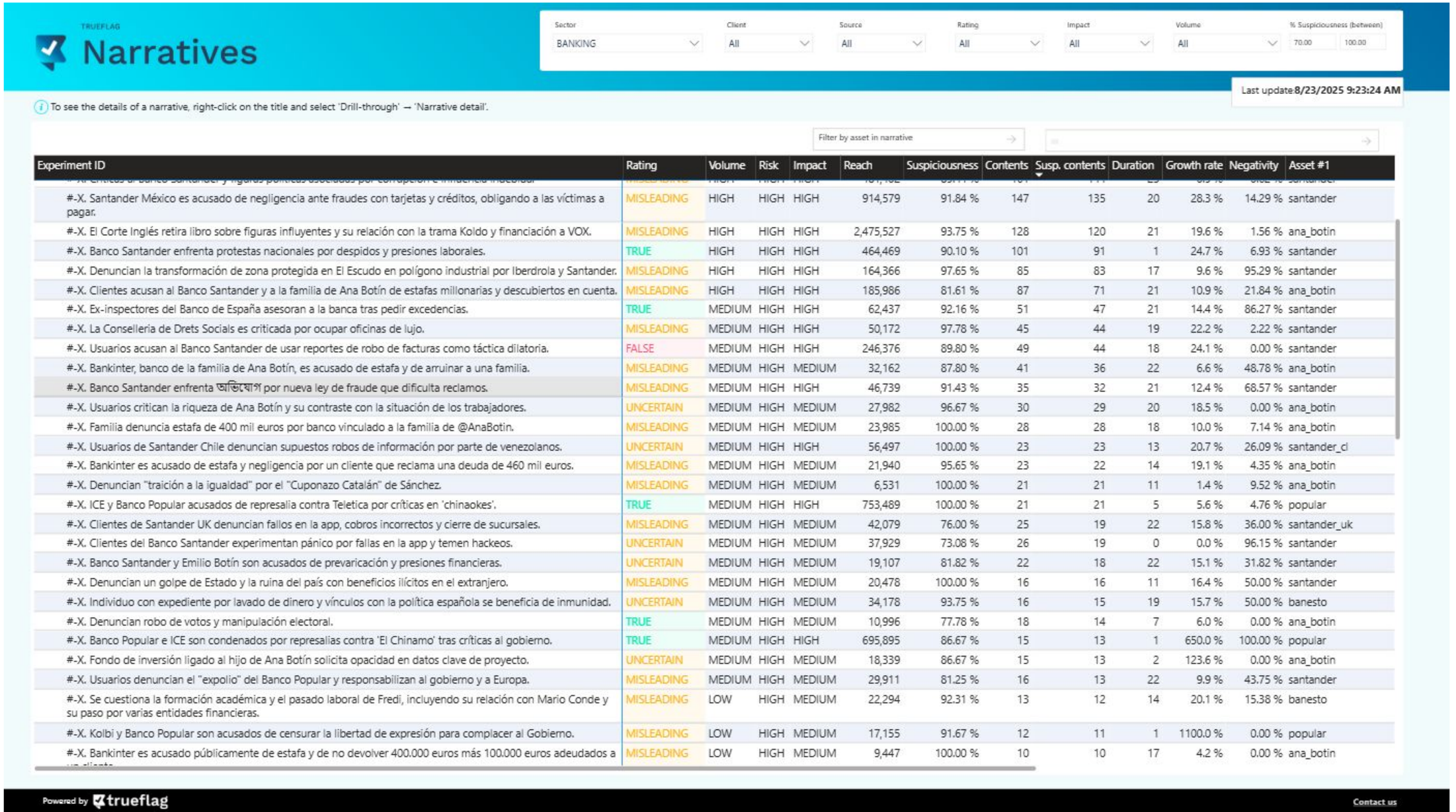
- **Security** requirements
- Evaluate potential automation of processes
- Integration and implementation requirements
- Workplan


trueflag
New narrative detection cycle per week
GPU
AI
Narrative Detection
Narrative Summarization
Verification Agent
7 days
Narrative List
targeted assets
1. keyword search
2. send data
X API
1 min
3. save
Data Adapter
3 min
raw data
MONITORING PROVIDER
4. start processing
Asset Labeler
5 min
5a. detection of toxicity
5b. assign to a narrative
5c. detection of fake news
15 min
Toxicity Detector
Narrative Prediction
1 day
Claim Matching
15 min
fact-checking articles
adverse narratives
Object of Attack
15 min
Disinfo Pattern Detection
15 min
DASHBOARD + ALERTS
1 hour
AI as a service (token consumption)
AI on own GPU servers

CLIENT DASHBOARD: MVP



CLIENT DASHBOARD: MVP



TRUEFLAG

Narratives

Sector: BANKING | Client: All | Source: All | Rating: All | Impact: All | Volume: All | % Suspiciousness (between): 70.00 – 100.00

Last update: 8/23/2025 9:23:24 AM

To see the details of a narrative, right-click on the title and select 'Drill-through' → 'Narrative detail'.

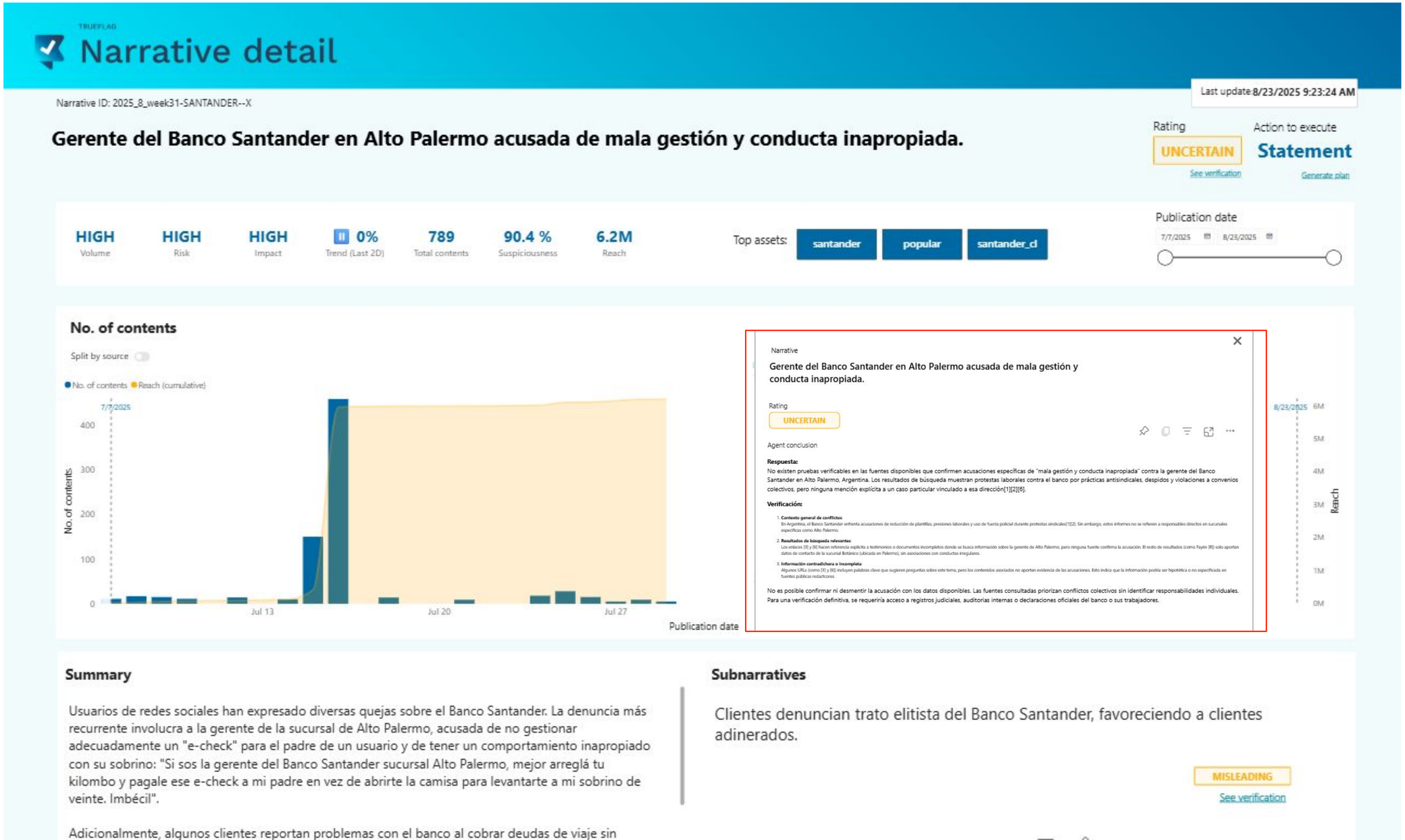
Filter by asset in narrative

Experiment ID	Rating	Volume	Risk	Impact	Reach	Suspiciousness	Contents	Susp. contents	Duration	Growth rate	Negativity	Asset #1
#-X. Santander México es acusado de negligencia ante fraudes con tarjetas y créditos, obligando a las víctimas a pagar.	MISLEADING	HIGH	HIGH	HIGH	914,579	91.84 %	147	135	20	28.3 %	14.29 %	santander
#-X. El Corte Inglés retira libro sobre figuras influyentes y su relación con la trama Koldo y financiación a VOX.	MISLEADING	HIGH	HIGH	HIGH	2,475,527	93.75 %	128	120	21	19.6 %	1.56 %	ana_botin
#-X. Banco Santander enfrenta protestas nacionales por despidos y presiones laborales.	TRUE	HIGH	HIGH	HIGH	464,469	90.10 %	101	91	1	24.7 %	6.93 %	santander
#-X. Denuncian la transformación de zona protegida en El Escudo en polígono industrial por Iberdrola y Santander.	MISLEADING	HIGH	HIGH	HIGH	164,366	97.65 %	85	83	17	9.6 %	95.29 %	santander
#-X. Clientes acusan al Banco Santander y a la familia de Ana Botín de estafas millonarias y descubiertos en cuenta.	MISLEADING	HIGH	HIGH	HIGH	185,986	81.61 %	87	71	21	10.9 %	21.84 %	ana_botin
#-X. Ex-inspectores del Banco de España asesoran a la banca tras pedir excedencias.	TRUE	MEDIUM	HIGH	HIGH	62,437	92.16 %	51	47	21	14.4 %	86.27 %	santander
#-X. La Conselleria de Drets Socials es criticada por ocupar oficinas de lujo.	MISLEADING	MEDIUM	HIGH	HIGH	50,172	97.78 %	45	44	19	22.2 %	2.22 %	santander
#-X. Usuarios acusan al Banco Santander de usar reportes de robo de facturas como táctica dilatoria.	FALSE	MEDIUM	HIGH	HIGH	246,376	89.80 %	49	44	18	24.1 %	0.00 %	santander
#-X. Bankinter, banco de la familia de Ana Botín, es acusado de estafa y de arruinar a una familia.	MISLEADING	MEDIUM	HIGH	MEDIUM	32,162	87.80 %	41	36	22	6.6 %	48.78 %	ana_botin
#-X. Banco Santander enfrenta অভিযোগ por nueva ley de fraude que dificulta reclamos.	MISLEADING	MEDIUM	HIGH	HIGH	46,739	91.43 %	35	32	21	12.4 %	68.57 %	santander
#-X. Usuarios critican la riqueza de Ana Botín y su contraste con la situación de los trabajadores.	UNCERTAIN	MEDIUM	HIGH	MEDIUM	27,982	96.67 %	30	29	20	18.5 %	0.00 %	ana_botin
#-X. Familia denuncia estafa de 400 mil euros por banco vinculado a la familia de @AnaBotin.	MISLEADING	MEDIUM	HIGH	MEDIUM	23,985	100.00 %	28	28	18	10.0 %	7.14 %	ana_botin
#-X. Usuarios de Santander Chile denuncian supuestos robos de información por parte de venezolanos.	UNCERTAIN	MEDIUM	HIGH	HIGH	56,497	100.00 %	23	23	13	20.7 %	26.09 %	santander_cl
#-X. Bankinter es acusado de estafa y negligencia por un cliente que reclama una deuda de 460 mil euros.	MISLEADING	MEDIUM	HIGH	MEDIUM	21,940	95.65 %	23	22	14	19.1 %	4.35 %	ana_botin
#-X. Denuncian "traición a la igualdad" por el "Cuponazo Catalán" de Sánchez.	MISLEADING	MEDIUM	HIGH	MEDIUM	6,531	100.00 %	21	21	11	1.4 %	9.52 %	ana_botin
#-X. ICE y Banco Popular acusados de represalia contra Teletica por críticas en 'chinaokes'.	TRUE	MEDIUM	HIGH	HIGH	753,489	100.00 %	21	21	5	5.6 %	4.76 %	popular
#-X. Clientes de Santander UK denuncian fallos en la app, cobros incorrectos y cierre de sucursales.	MISLEADING	MEDIUM	HIGH	MEDIUM	42,079	76.00 %	25	19	22	15.8 %	36.00 %	santander_uk
#-X. Clientes del Banco Santander experimentan pánico por fallas en la app y temen hackeos.	UNCERTAIN	MEDIUM	HIGH	MEDIUM	37,929	73.08 %	26	19	0	0.0 %	96.15 %	santander
#-X. Banco Santander y Emilio Botín son acusados de prevaricación y presiones financieras.	UNCERTAIN	MEDIUM	HIGH	MEDIUM	19,107	81.82 %	22	18	22	15.1 %	31.82 %	santander
#-X. Denuncian un golpe de Estado y la ruina del país con beneficios ilícitos en el extranjero.	MISLEADING	MEDIUM	HIGH	MEDIUM	20,478	100.00 %	16	16	11	16.4 %	50.00 %	santander
#-X. Individuo con expediente por lavado de dinero y vínculos con la política española se beneficia de inmunidad.	UNCERTAIN	MEDIUM	HIGH	MEDIUM	34,178	93.75 %	16	15	19	15.7 %	50.00 %	banesto
#-X. Denuncian robo de votos y manipulación electoral.	TRUE	MEDIUM	HIGH	MEDIUM	10,996	77.78 %	18	14	7	6.0 %	0.00 %	ana_botin
#-X. Banco Popular e ICE son condenados por represalias contra 'El Chinamo' tras críticas al gobierno.	TRUE	MEDIUM	HIGH	HIGH	695,895	86.67 %	15	13	1	650.0 %	100.00 %	popular
#-X. Fondo de inversión ligado al hijo de Ana Botín solicita opacidad en datos clave de proyecto.	UNCERTAIN	MEDIUM	HIGH	MEDIUM	18,339	86.67 %	15	13	2	123.6 %	0.00 %	ana_botin
#-X. Usuarios denuncian el "expolio" del Banco Popular y responsabilizan al gobierno y a Europa.	MISLEADING	MEDIUM	HIGH	MEDIUM	29,911	81.25 %	16	13	22	9.9 %	43.75 %	santander
#-X. Se cuestiona la formación académica y el pasado laboral de Fredi, incluyendo su relación con Mario Conde y su paso por varias entidades financieras.	MISLEADING	LOW	HIGH	MEDIUM	22,294	92.31 %	13	12	14	20.1 %	15.38 %	banesto
#-X. Kolbi y Banco Popular son acusados de censurar la libertad de expresión para complacer al Gobierno.	MISLEADING	LOW	HIGH	MEDIUM	17,155	91.67 %	12	11	1	1100.0 %	0.00 %	popular
#-X. Bankinter es acusado públicamente de estafa y de no devolver 400.000 euros más 100.000 euros adeudados a un cliente	MISLEADING	LOW	HIGH	MEDIUM	9,447	100.00 %	10	10	17	4.2 %	0.00 %	ana_botin

Powered by trueflag

Contact us

CLIENT DASHBOARD: MVP



KEY DATA

1M posts per week

4 networks monitored: X, TG, TikTok, Digital Press

In 21st century, something can make conversation impossible: if trust between people collapses. The main space to converse, online, is flooded by non-human entities that pretend to be human beings. You read a text, hear an audio, see a video, and you don't know if it's real.

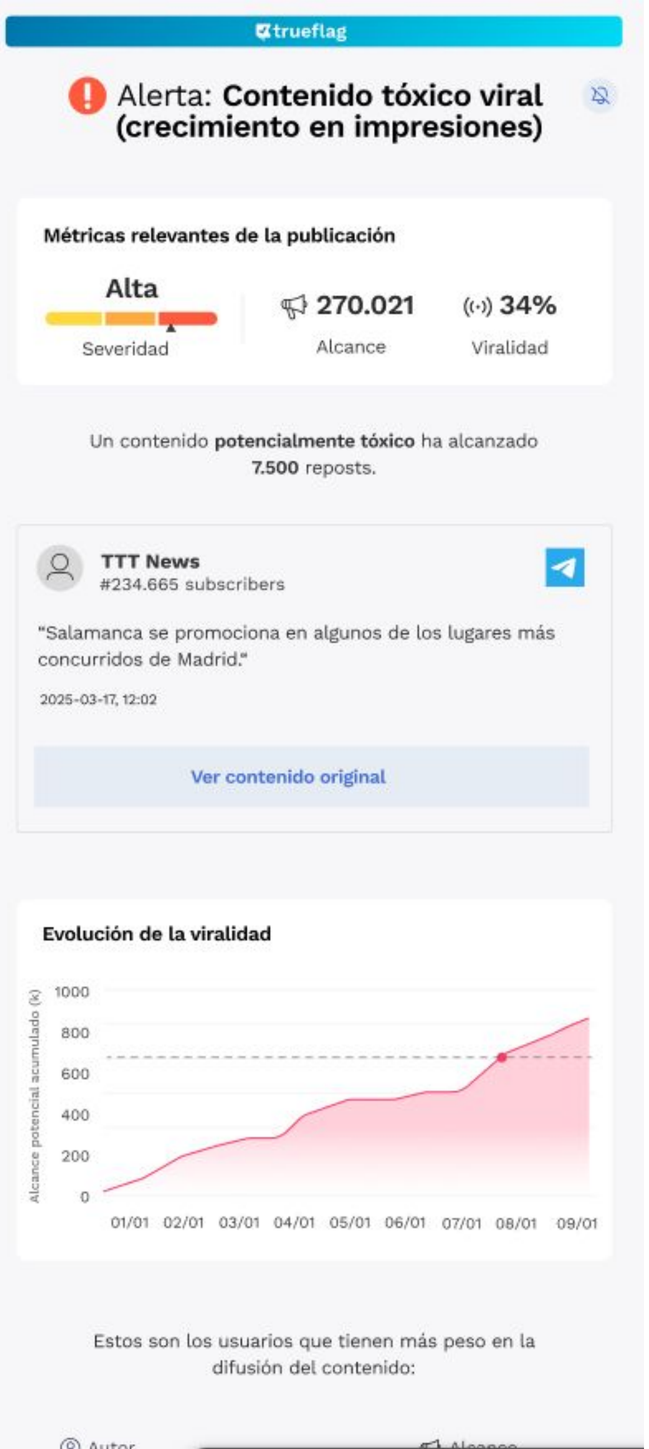
ALERT ENGINE

Clients need pre-emptive alerts of potential crisis

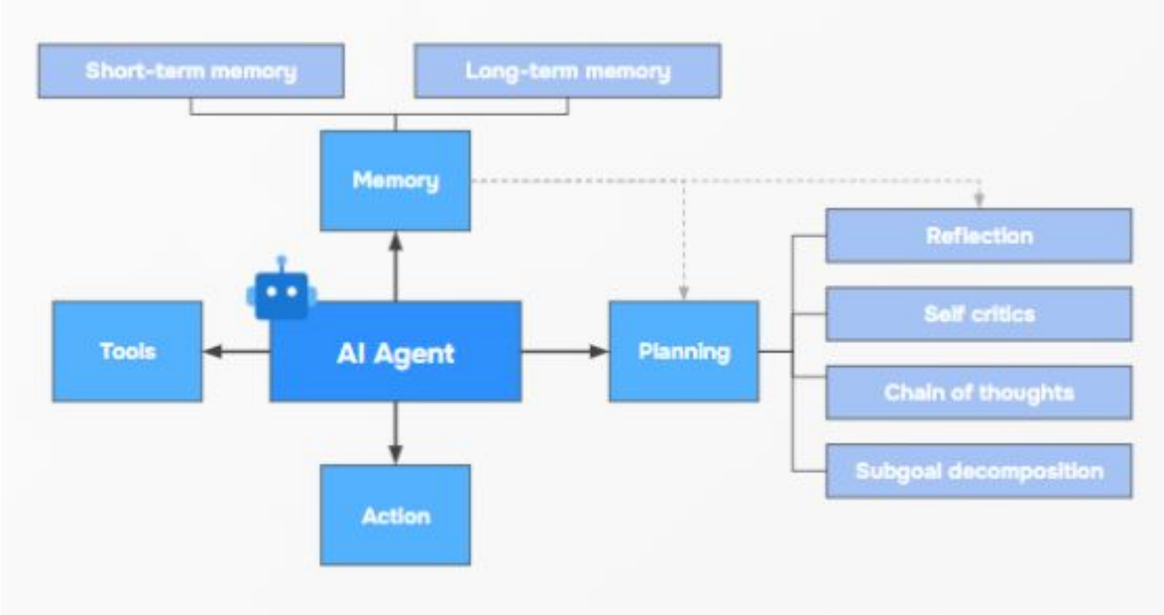
Severity levels and contextual information must be provided for easy prioritization by Communication departments -> Prioritization AI agent

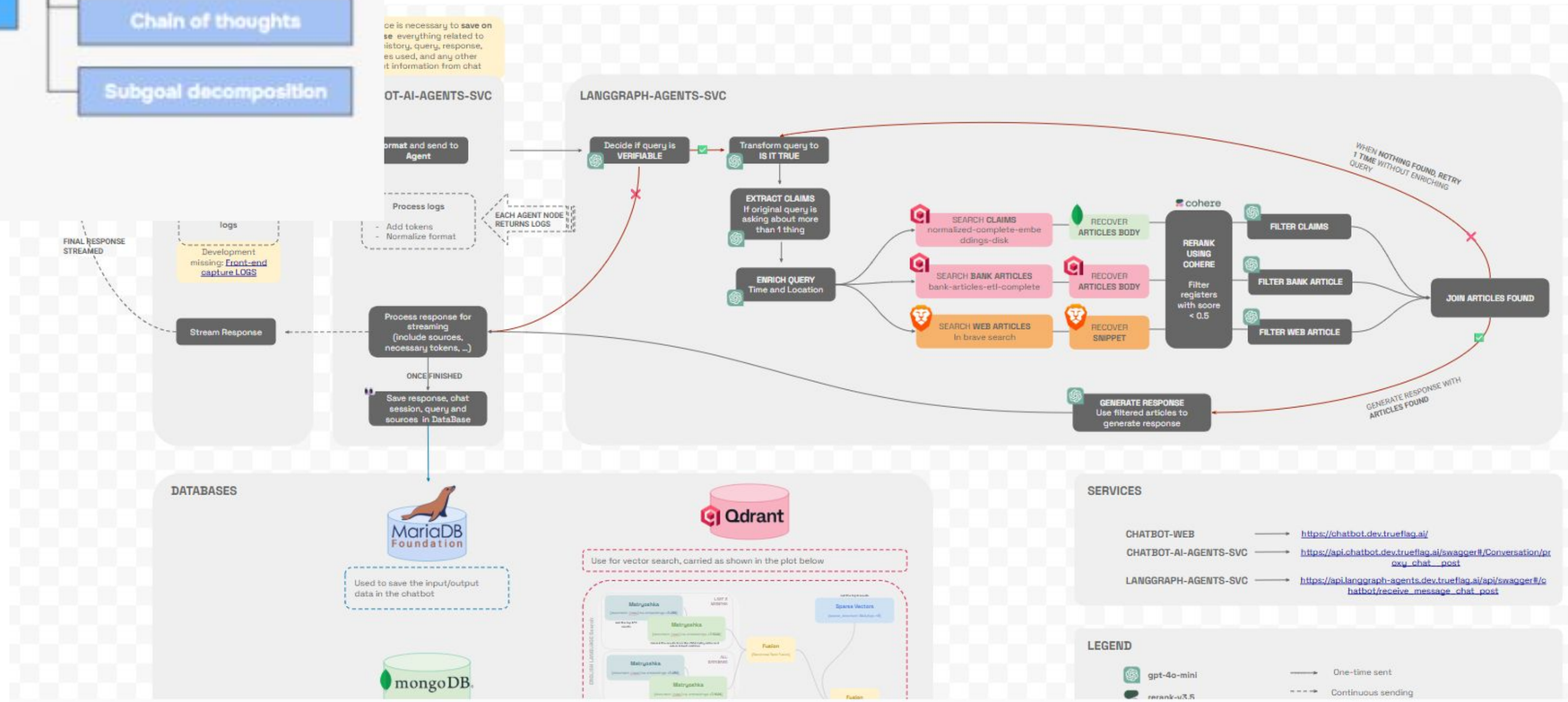
Dashboard for crisis management or broader analysis of current situation

Highly interested in: Multimedia Content



AI VERIFICATION AGENT





AI VIDEO AGENT

MULTIMODAL — Asset Labeler
- Brand mentions on video images
- Brand mentions on audio transcript
- Brand mentions on shown documents

MULTIMODAL — Claim Extraction
- Summarize main message in one single claim

MULTIMODAL — Summarize video content
- Automated short description of what is being stated in the video

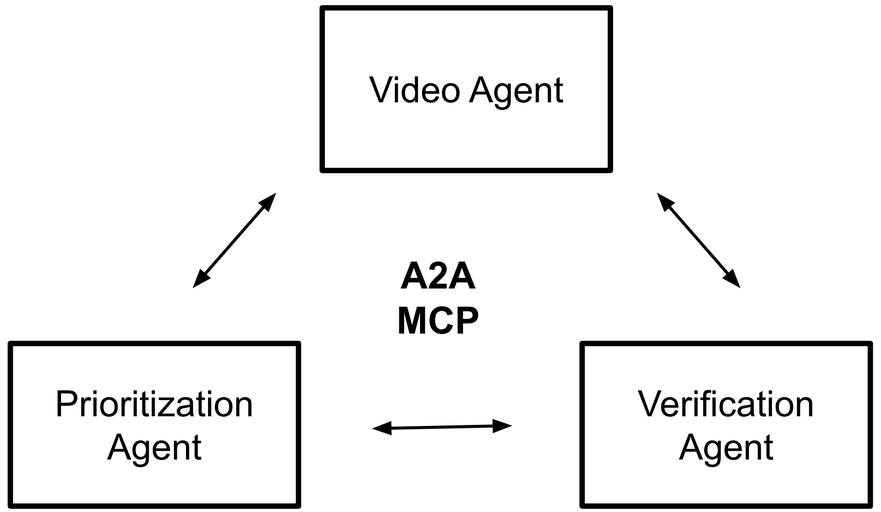
MULTIMODAL — Attack Detection
- Identify targeted brand/products
- Identify tone, contextual info, human references, mixed content
- Extract homogeneous claims for semantic clusterization

Regular Workflow

WHICH ACCOUNTS TO TRACK?

HOW TO MANAGE COMPUTATIONAL COSTS?

HOW TO MAKE A MULTIPLATFORM ANALYSIS OF THE CONTENT?



Achieved!

Grey Web

- Automatic Expansion Algorithms on Telegram
- Automatic detection algorithms for suspicious domains

Automated verification

- Narrative Verification Module
- Verification Chatbots

Algorithms focused on individual content detection

- Toxicity Detection
- Disinformation Patterns detection
- Claim matching
- Claim detection
- Adverse narratives module

Narrative extraction

- Narrative Identification Module
- Narrative Contextualization Module
- TikTok Content Analysis Module

Grey Web

Automatic Expansion Algorithms for TikTok, X and for domains

Narrative extraction

Risk Prioritization Module

Automated verification

- Integrate deepfake detection solutions
- Improving reasoning and analysis of multimodal or domain-specific information

AI MODEL LIST

NLI models

https://huggingface.co/Newtral/mDeBERTa-v3-base-nli-media-v0.2
https://huggingface.co/Newtral/DeBERTa-v3-large-nli-media-v0.1
microsoft/mdeberta-v3-base, microsoft/Multilingual-MiniLM-L12-H384
https://huggingface.co/Newtral/deberta-v3-base-zeroshot-v1.1-all-33-industry-classification-half-freeze
https://huggingface.co/Newtral/Multilingual-MiniLM-L12-H384-industry-classification-no-freeze

Reranker

https://huggingface.co/Newtral/Qwen3-Reranker-0.6B-cm-reranking
BAAI/bge-m3
BAAI/bge-reranker-v2-gemma

Embeddings

jinaai/jina-embeddings-v3
BAAI/bge-m3

Disinfo Patters

LLM (SFT instruct + preference aligned) right arrow short reasoning (without RL)
QWEN-4B-bfloat16
https://huggingface.co/Newtral/Qwen3-4B-kto-disinformation-bfloat16-v4
QWEN-4B-GGUF (8 bit local quantization)
https://huggingface.co/Newtral/Qwen3-4B-kto-disinformation-v4-GGUF
QWEN-4B -AWQ (4 bit post training quantization → current vLLM deploy)
https://huggingface.co/Newtral/Qwen3-4B-kto-disinformation-AWQ-v4

Transformer-encoder (bert-like, multilabel sequence classification)
https://huggingface.co/Newtral/EuroBERT-610m-multilabel-generalist-disinformation

Closed source LLMs

OpenAI gpt4o-mini
OpenAI gpt4.1-mini
gemini-flash-2.0
OpenAI o3-mini

ONGOING PROJECTS

IFCN GLOBAL CHATBOT

A white-brand chatbot for any fact-checking website

- Tech leaders of the Global Chatbot Development
- Global collaborative database of fact-checks
- Conversational agent for general public

EFCSN REPOSITORY

European repository of fact-checks

- MCP tool to support LLM monetization of data repository
- Directory of trusted/untrusted sources
- Integration with EDMO hubs

Mobile World Congress / Google.org

Claim Detection / Claim Matching API

- Automated fact-checking platform for any Spanish/Latam fact-checker

FACTFLOW

FactFlow

Search messages...

Logout

Post Frequency

last 3 Months

Published messages

398,181

last 3 Months

Suspicious Content

37,665

last 3 Months

All Content

398,181 messages from 8,011 Channels

All languages | last 3 Months | Recent First | Reset All Filters

NM

0 0

Classify content

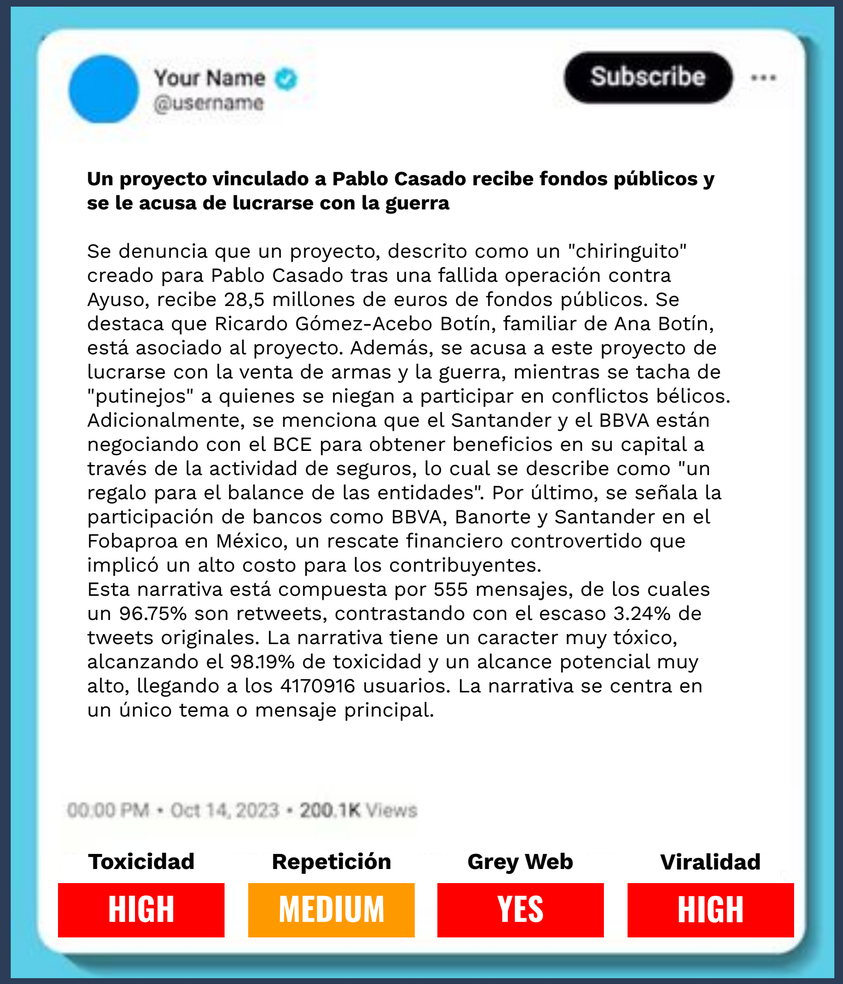
El payaso judío Kirk hace seis meses: "¿Quién se beneficia de la guerra? El complejo militar-industrial y los oligarcas de la clase dirigente de Ucrania".

OPEN CHALLENGES:
Defensive AI: assistant to fight back disinformation

trueflag

who do you trust?

ACTIONABLE INSIGHTS



Narrative risk KPIs

AI

Automated Risk Evaluation

Nivel de Riesgo: VERY HIGH

Justificación

- La narrativa vincula personajes políticos con “assets” monitorizados por TrueFlag (miembros de la familia Botín)
- La narrativa trata un tema sensible (financiación guerra) en un contexto europeo bélico en el que se está discutiendo el rearme de Europa.
- Se han identificado otras tres narrativas vinculando al Banco Santander con la financiación de armas. Podría formar parte de una campaña más amplia orientada a dañar la marca.
- Varios de los contenidos tienen un tono altamente emocional lo que puede influenciar al público y facilitar su propagación
- El alcance actual de la narrativa es actualmente alto (+4M usuarios)

CONSECUENCIAS DE IGNORAR LA NARRATIVA

La narrativa tiene un potencial muy alto de amplificación y puede formar parte de una campaña planificada. Una escalada sin seguimiento sería muy peligrosa, pudiendo dañar la reputación de la empresa entre amplios colectivos sensibilizados con la temática.

Recomendación

ACT

Dada la combinación de alcance, temática sensible y nivel de eco de la historia en redes se recomienda preparar un plan de respuesta temprana contra la narrativa.

DEFINE PLAN

ACTIONABLE INSIGHTS

Propuestas de Acción

STEP 1

ACCIÓN 1

Implementar un seguimiento intensivo en medios y redes sociales para detectar cambios en el tono y frecuencia de publicaciones. Establecer alertas programadas sobre hashtags como #BancaArmada.

ACCIÓN 2

Preparar plantillas de respuesta basadas en datos verificables para uso del equipo de comunicación en caso de que la narrativa se amplificara.

ACCIÓN 3

Coordinar con relaciones públicas y comunicación interna para revisar y ajustar, si es necesario, la estrategia de respuesta en tiempo real.

ACCIÓN 4

Propón otras acciones que quieras incluir al plan

NEXT

TrueFlag's AI proposes several actions to fight back the narrative

Customize the plan
(human in the loop approach)

Líneas Generales del Plan

STEP 2

Análisis

ESTADO ACTUAL

La narrativa ha alcanzado una tracción considerable. Incluye contenido emocional y se han identificado elementos de la Grey Web en su transmisión. Existe un riesgo alto de viralización y potencial paso a la prensa digital.

DESAFÍOS Y OPORTUNIDADES

La narrativa ha alcanzado una tracción considerable. Incluye contenido emocional y se han identificado elementos de la Grey Web en su transmisión. Existe un riesgo alto de viralización en X y Telegram. Potencial impacto en prensa digital.

Plan de acción

FASES

1. **Fase de Inicio**:
   - Monitorización intensiva en redes para identificar cambios en la narrativa
   - Elaboración y difusión de mensajes preliminares que establezcan el posicionamiento de la entidad frente a la financiación de acciones bélicas
2. **Fase de Desarrollo**:
   - Implementación de plantillas de respuesta y coordinación entre equipos de comunicación, relaciones públicas y crisis.
   - Ejecución de acciones interactivas para responder a comentarios y resolver dudas en tiempo real.
3. **Fase de Consolidación**:
   - Evaluación de resultados, análisis de KPIs y ajuste de estrategias.
   - Comunicación interna continua y actualización de protocolos según la evolución de la narrativa.

ROLES

Equipo multidisciplinar:

- **Equipo de Monitoreo y Análisis Digital**: responsable de la vigilancia de redes y análisis de sentimiento
- **Equipo de Comunicación y Crisis**: encargado de preparar y difundir comunicados, gestionar respuestas y coordinar con relaciones públicas
- **Equipo de Relaciones Internas**: que garantice la coherencia del mensaje y el alineamiento de toda la organización.
- **Soporte Técnico y Analítico**: para el uso de herramientas de monitorización y análisis de datos.
- **Consultores externos en Desinformación**: como soporte en la revisión de los desmentidos a publicar

MODIFY PLAN

APPROVE PLAN

AI generates a concrete plan based on this main guidelines with specific tasks and actions for each stakeholder

OPEN CHALLENGES:

- Multi-agent architecture
- Multimodal disinformation (out of context, deepfakes)
- Grey Web automated expansion
- Cost efficiency

trueflag

who do you trust?

ruben.miguez@trueflag.ai