Harnessing data to combat Propaganda and Disinformation

Nesin Veli,

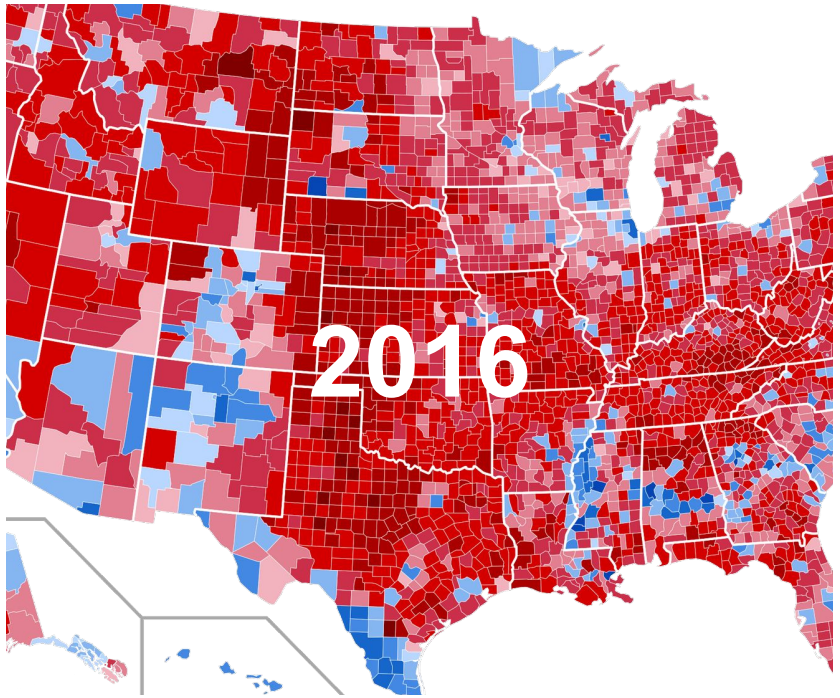
Products & Services Manager @ Identrics

Web Data Extraction Summit | October 2023

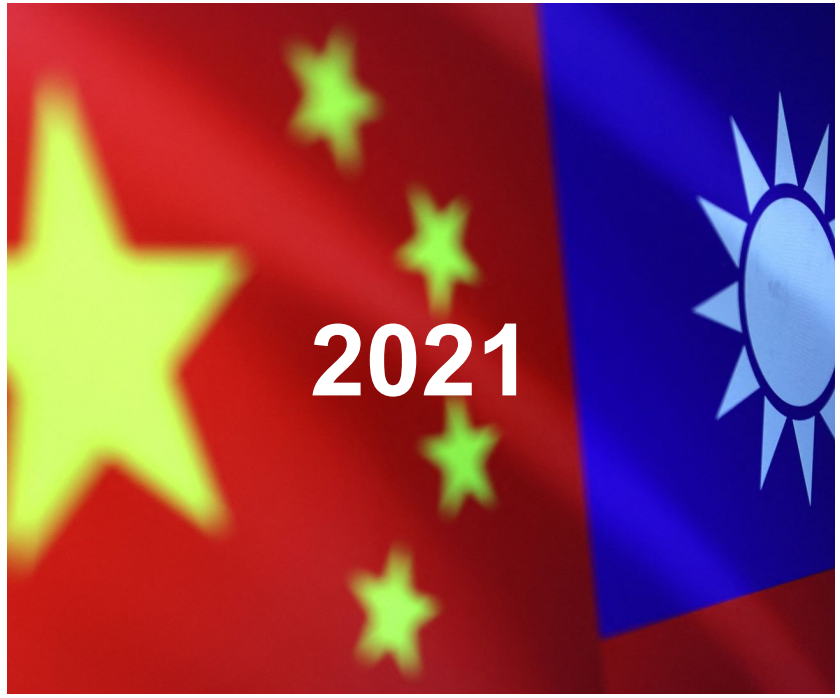
Cognitive warfare

identrics


Sea
Azo
CRIMEA
Kerch
Simferopol
evastopol
Bahkchysaray
2014


2016

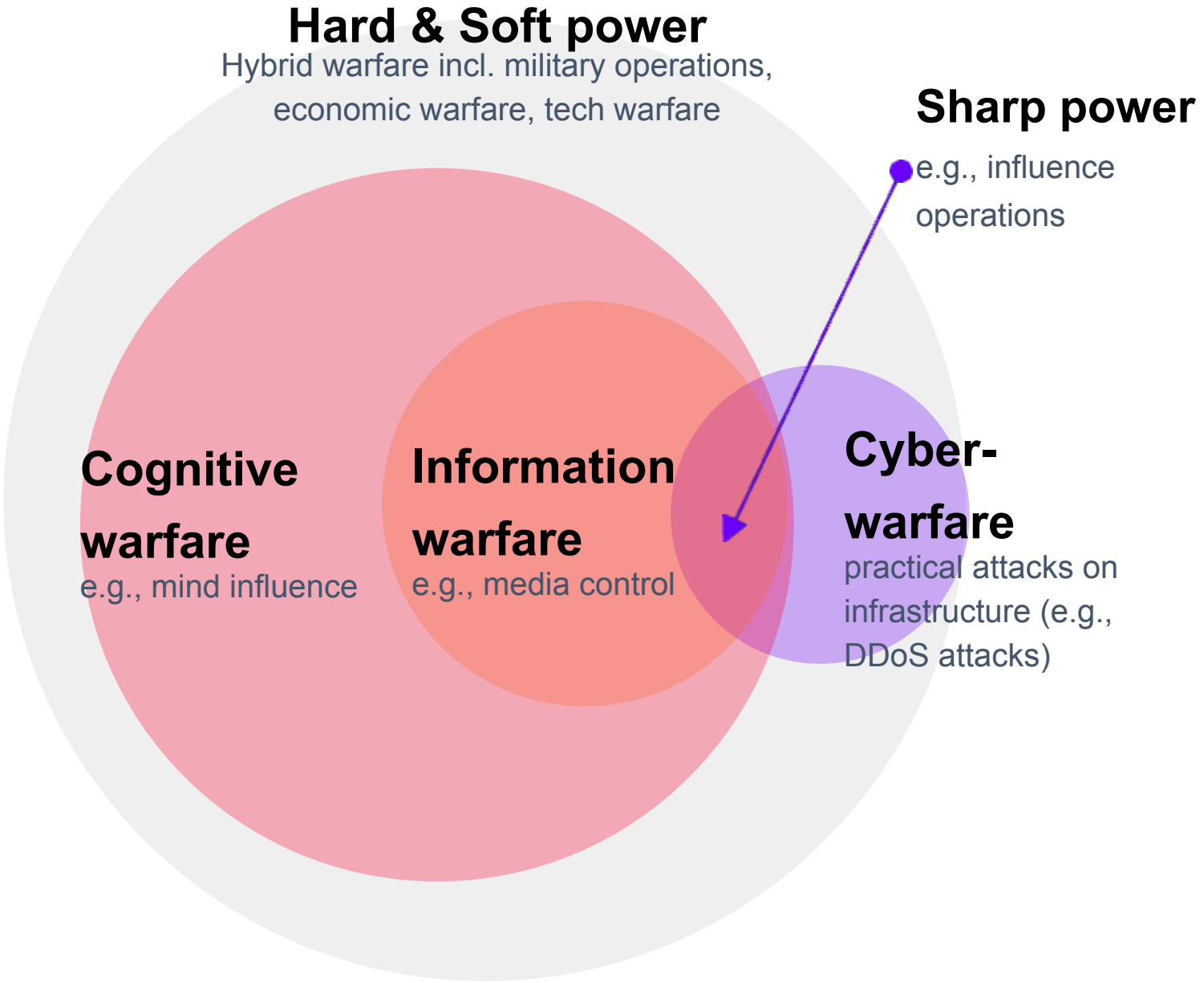

2020


2021

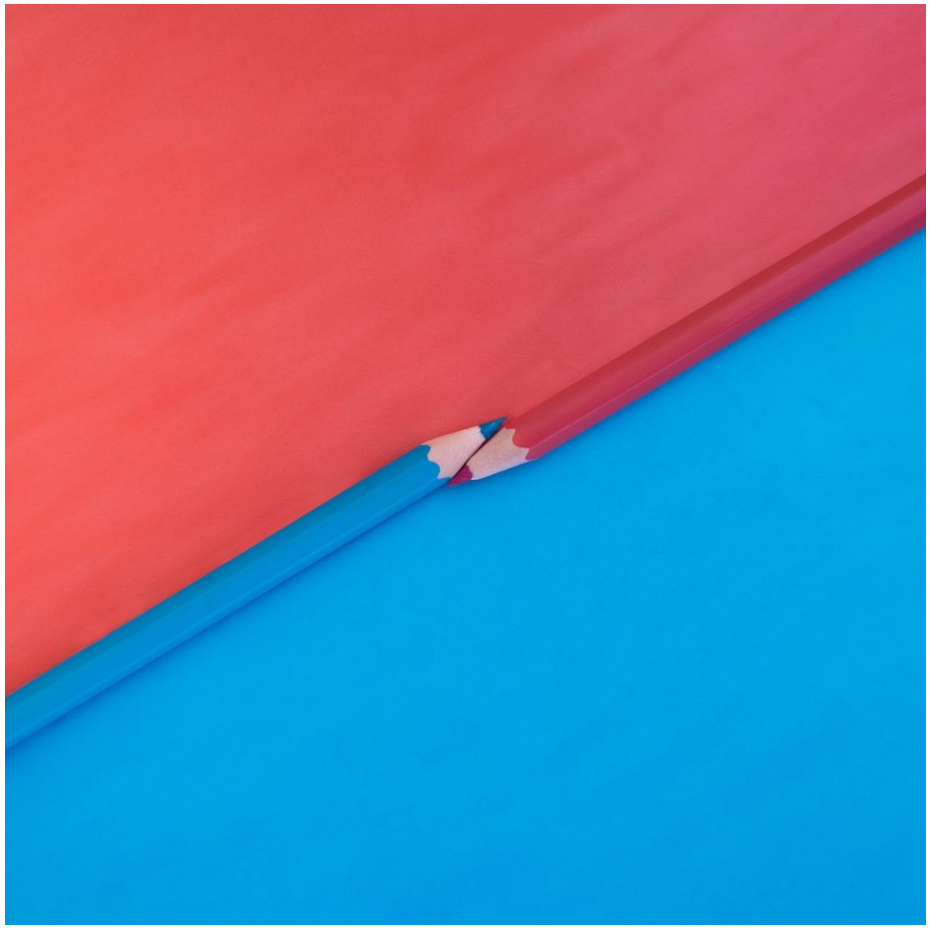

2022


2023

Cognitive warfare



Cognitive warfare



Division and distrust

societal split

government distrust

enable saviour figures

Cognitive warfare

Hearts and minds

it is not just about facts and lies

it is about angst

identrics

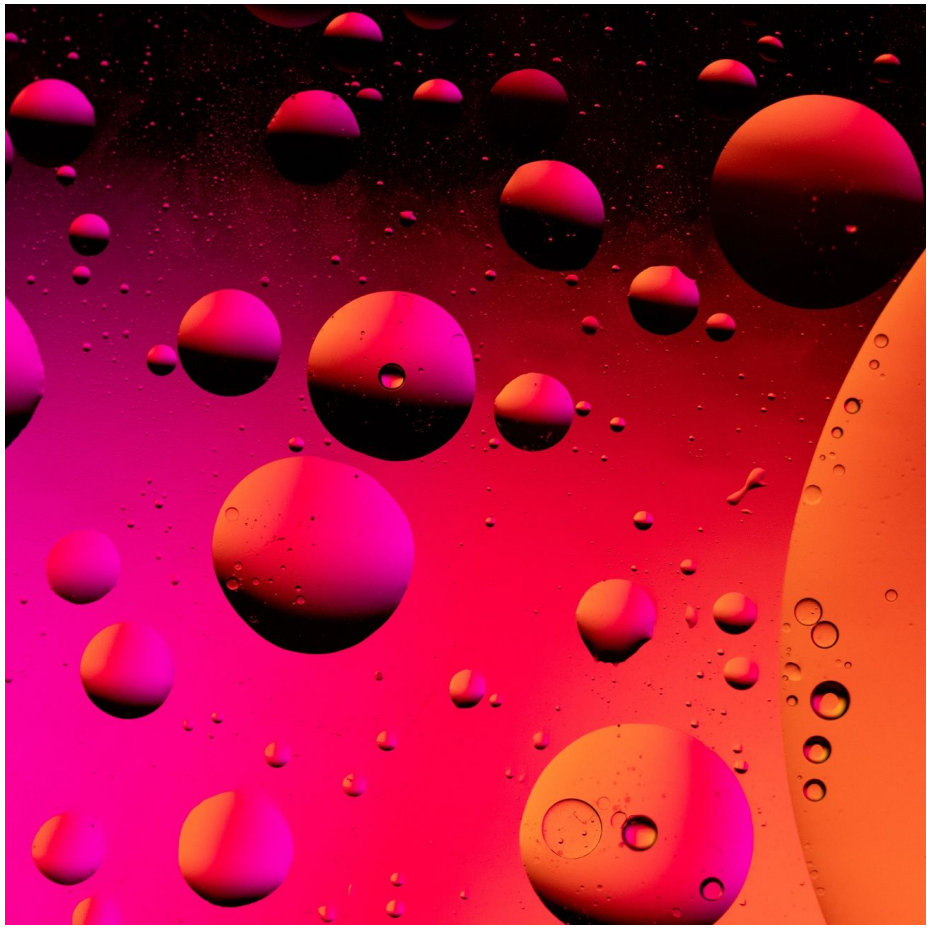
Cognitive warfare

Actors

rise of non-state actors

identrics

Cognitive warfare



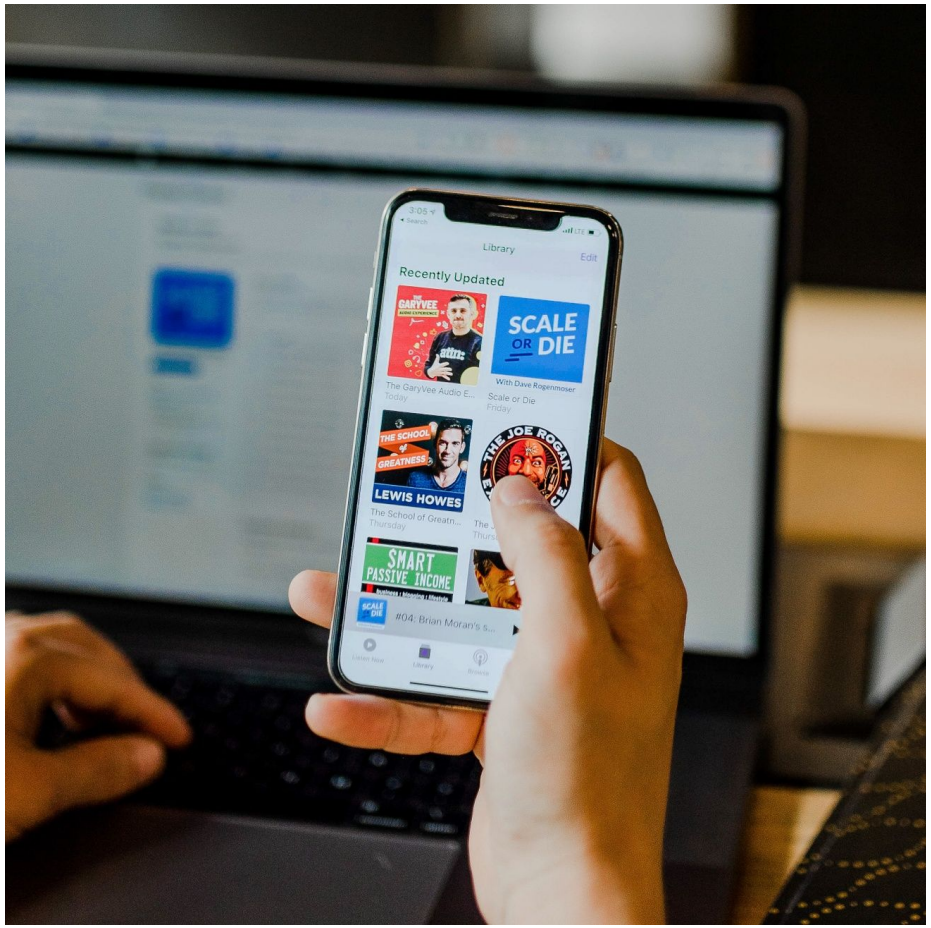
Comfort info bubbles

algorithm-pushed news calcified info bubbles

targets are more easily agitated

identrics

Cognitive warfare



Media literacy

ideal and too slow

identrics

Cognitive warfare

Tech resilience

automation enabling human decision

monitoring and alerting systems

identrics

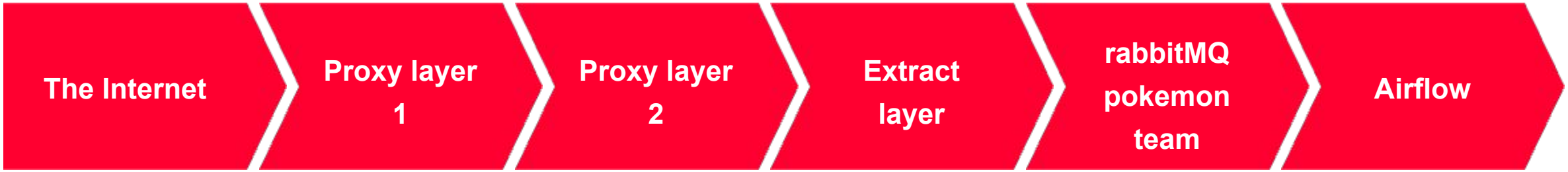

identrics

Data extraction

identrics

Extraction ecosystem

Encompass all aggregated data including disinfo monitoring



- Mainstream media
- Fringe media
- Fast data

<10% of unstructured data traffic
>90% of structured data traffic

Residential for some use cases

- Multiple Scrapy instances
- Kraken source GUI

- Empoleon
- Ludicolo
- Pikachu
- Lapras
- Caterpie
- Lunala
- Snorlax

Three lonely workers

Extraction ecosystem cont.

Encompass all aggregated data including disinfo monitoring

- Elastic
- Logstash
- Kibana

- Engagement
- Topic
- Entity
- Sentiment
- Knowledge base

- Hate speech alerts
- Source credibility profiles
- HITL model training
- Clean training data

Data per use case

Fast data

- new social media
- mainstream news user comments

"Fringe" media

- mimic mainstream design
- infect the mainstream
- replace the mainstream

Mainstream

- prone to misinformation
- prone to narrative hijack
- prone to algorithm mimicry

Narrative tracking

Origin to end point

Insights

Actionable alerting

Control vs censorship

identrics

Disinfo playbook

identrics

Propaganda, disinfo, hate speech

Propaganda

- **Broad, general** messages
- **No fact checking**

Disinformation

- Manipulated **content**
- Manupilated **context**

Malinformation

- Direct **cyber attacks**
- **Infomration leaks**

Misinformation

- **Unaware actors**
- **Main way of dissemenation**

Hate speech

- **Marginalizing** groups
- Media **narrative hijack**
- Media **regulatory fines**

identrics

Natural emotional language

Loaded language

Outrage as Donald Trump suggests injecting disinfectant to kill virus.

Name calling

WHO: Coronavirus emergency is '**Public Enemy Number 1**'.

Repetition

I still have a **dream**. It is a dream deeply rooted in the American **dream**. I have a **dream** that one day . . .

Appeal to fear

A dark, impenetrable and "irreversible" winter of persecution of the faithful by their own shepherds will fall.

Exaggeration or minimization

Coronavirus '**risk to the American people remains very low**', Trump said.

Doubt

Can the same be said for the Obama Administration?

Slogans

"BUILD THE WALL!" Trump tweeted.

Bandwagon

He tweeted, "**EU no longer considers #Hamas a terrorist group. Time for US to do same.**"

identrics

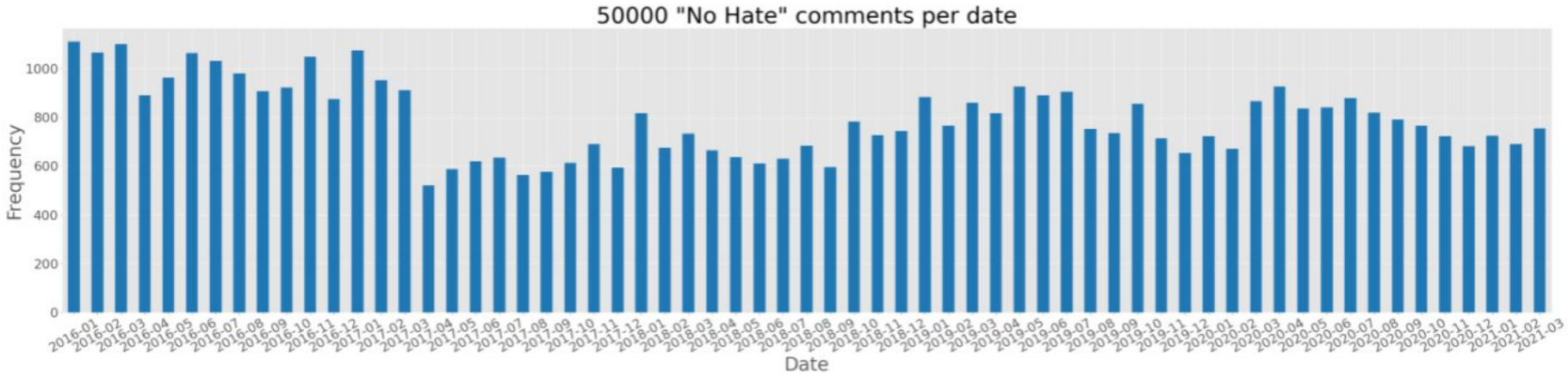
Hate speech alerting system

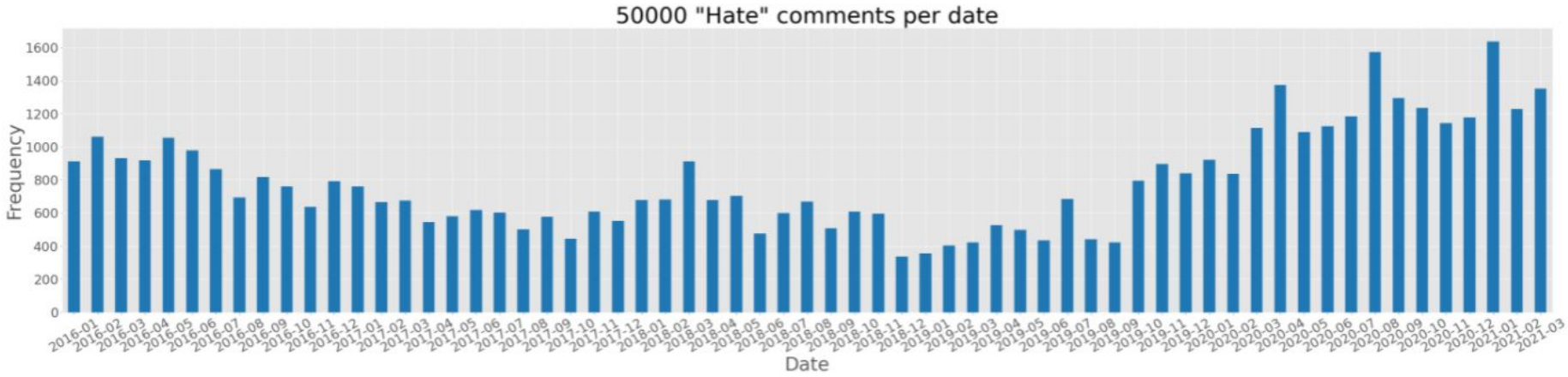
identrics

Hate speech raw dataset

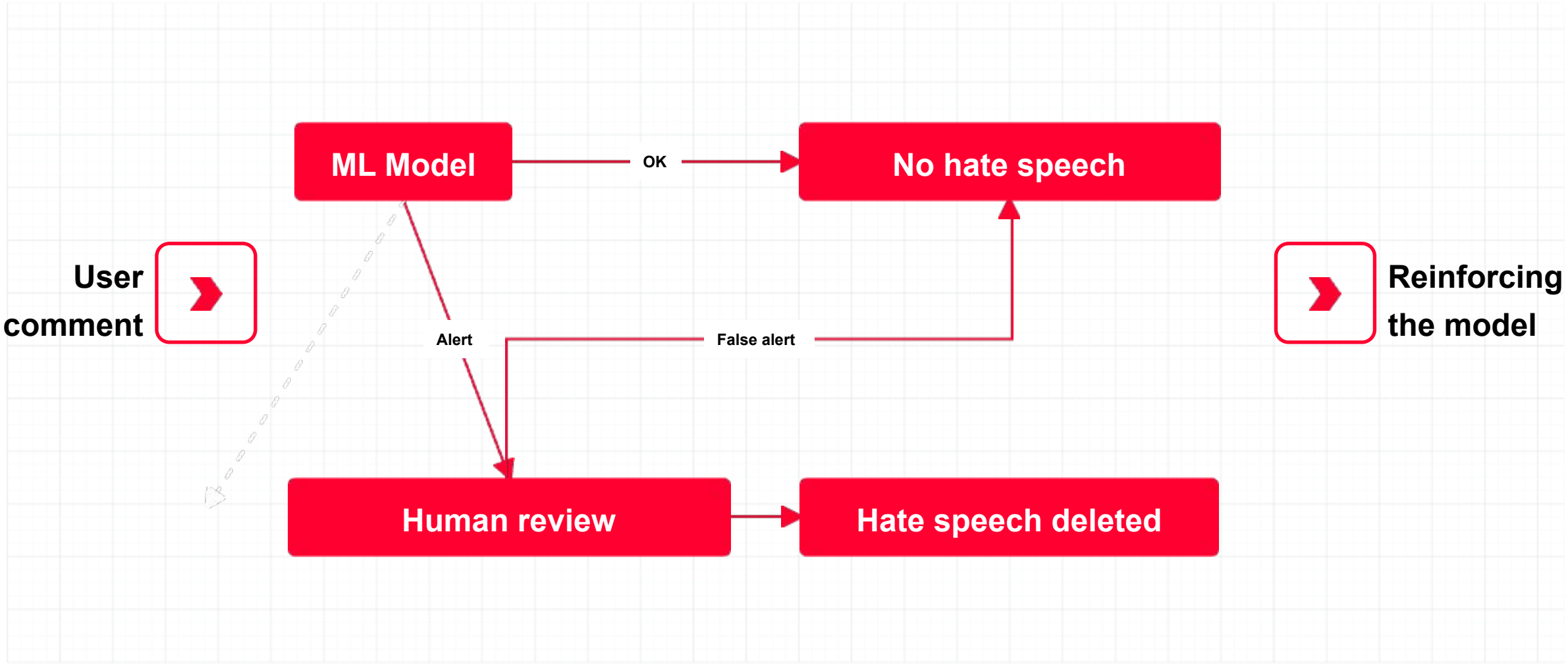
Moderator team annotation	Comment count
No hate speech	**4,000,000**
Hate speech	**150,000**
Adverts and spam	90,000
Different language	12,000
Political campaign	300
Not GDPR compliant	120

No hate vs hate speech histogram





Hate speech model reinforcement



Neural network results

Corpus: 419306 train + 52414 dev + 52413 test

```
Results:
- F-score (micro) 0.7993
- F-score (macro) 0.7951
- Accuracy 0.7993

By class:
              precision    recall  f1-score   support

          ok     0.8150    0.8339    0.8243     29596
        hate     0.7779    0.7544    0.7660     22817

   micro avg     0.7993    0.7993    0.7993     52413
   macro avg     0.7964    0.7942    0.7951     52413
weighted avg     0.7988    0.7993    0.7989     52413
 samples avg     0.7993    0.7993    0.7993     52413
```

Source credibility profiles

identrics

Outlet and author credibility

Original content

document clustering tracks copied content across sources

No info (clickbait) titles

recognition model

Topic dispersion

topic annotation

Staff transparency

author names, about us, editorial

Ownership transparency

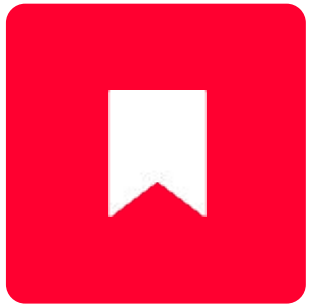
owner, publisher, PEP connections

Author profile

all of the above

identrics

Campaign pattern data



Topic dispersion

topic annotation

Komsomolskaya Pravda

on the eve of the Crimea annexation
the outlet starts publishing
exclusively sports content

Red flags

outlet and author profiles
along with historical patterns
are used to train recognition models

identrics

The five questions of disinfo campaigns

- Is there an active campaign?
- What is the origin point and current spread?
- Who are the actors?
- What are the goals?
- Is the pattern new or known?

identrics

LLMs

identrics

LLM and disinformation

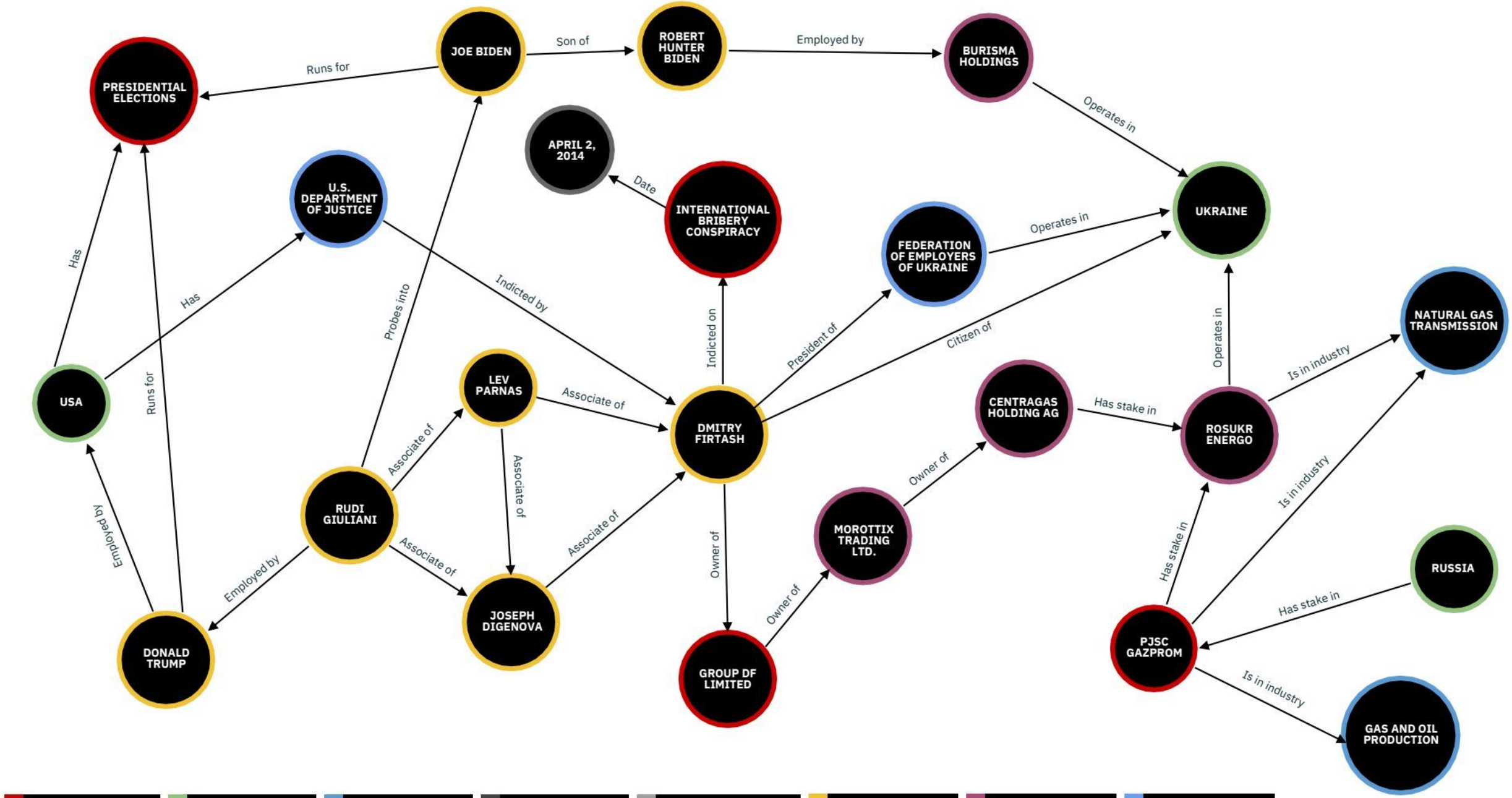
Threats

- Synthetic disinfo data fed back to the model
- Personalised disinfo
- Enable new bad actors
- Boost bot/cyborg networks
- Deep fakes and composite pictures

Opportunities

- Unstructured to structured data extraction
- Source credibility profiles
- Campaign pattern tracking
- Supervised fact-checking
- Cross border narrative tracking

identrics


PRESIDENTIAL ELECTIONS
JOE BIDEN
Son of
ROBERT HUNTER BIDEN
Employed by
BURISMA HOLDINGS
Runs for
Operates in
APRIL 2, 2014
Date
INTERNATIONAL BRIBERY CONSPIRACY
U.S. DEPARTMENT OF JUSTICE
FEDERATION OF EMPLOYERS OF UKRAINE
Operates in
UKRAINE
Has
Has
Indicted by
Probes into
Indicted on
President of
Citizen of
Operates in
NATURAL GAS TRANSMISSION
Is in industry
USA
Runs for
LEV PARNAS
Associate of
DMITRY FIRTASH
CENTRAGAS HOLDING AG
Has stake in
ROSUKR ENERGO
Associate of
Associate of
Associate of
Owner of
Owner of
Is in industry
Employed by
RUDI GIULIANI
Associate of
Associate of
MOROTTIX TRADING LTD.
Owner of
Has stake in
RUSSIA
Employed by
Associate of
Owner of
Has stake in
DONALD TRUMP
JOSEPH DIGENOVA
GROUP DF LIMITED
PJSC GAZPROM
Is in industry
GAS AND OIL PRODUCTION
NEWS/EVENTS
COUNTRIES
INDUSTRIES
TIMESTAMPS
RELATIONSHIPS
PEOPLE
COMPANIES
INSTITUTIONS

Custom models

Focused

Disinfo specific

Bad actor modus operandi

Clean

Remove known hate speech

No synthetic disinfo contamination

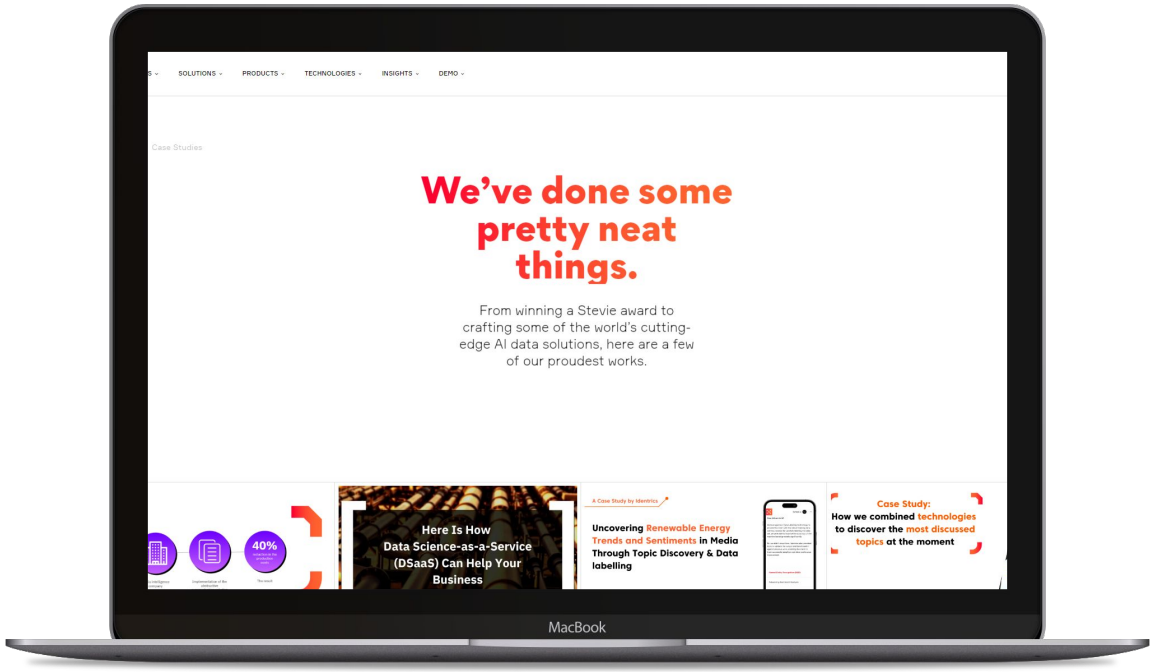
Use case specific

Hate speech per outlet

Manageable corpora

identrics

Useful links



- Identrics case studies
- Identrics blog
- Bellingcat Online Investigation Toolkit
- The Code of Practice on Disinformation

identrics

Citations

"How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars"

"SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles"

identrics

THANK YOU

nesin.veli@identrics.ai

www.identrics.ai

identrics