Lokale AI voor de burger

30/1/2024

AGENTSCHAP
BINNENLANDS
BESTUUR

DIGITAAL
VLAANDEREN

vvsg

Duiding van zorgwekkende fenomenen online met de hulp van AI

Gijs van Beek, Textgain

textgain

deDUIDER

textgain

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

AMiCA

textgain

PROJECT GREY

factcheck.
vlaanderen

DTCT
DETECT THEN ACT

RHETORIC

IMSyPP

COMMIT

EOOH
EUROPEAN OBSERVATORY OF ONLINE HATE

textgain
academy

BENEDMO

1995 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021

Social Media Monitoring

De Standaard Meest recent Binnenland Coronacrisis Meer ▾

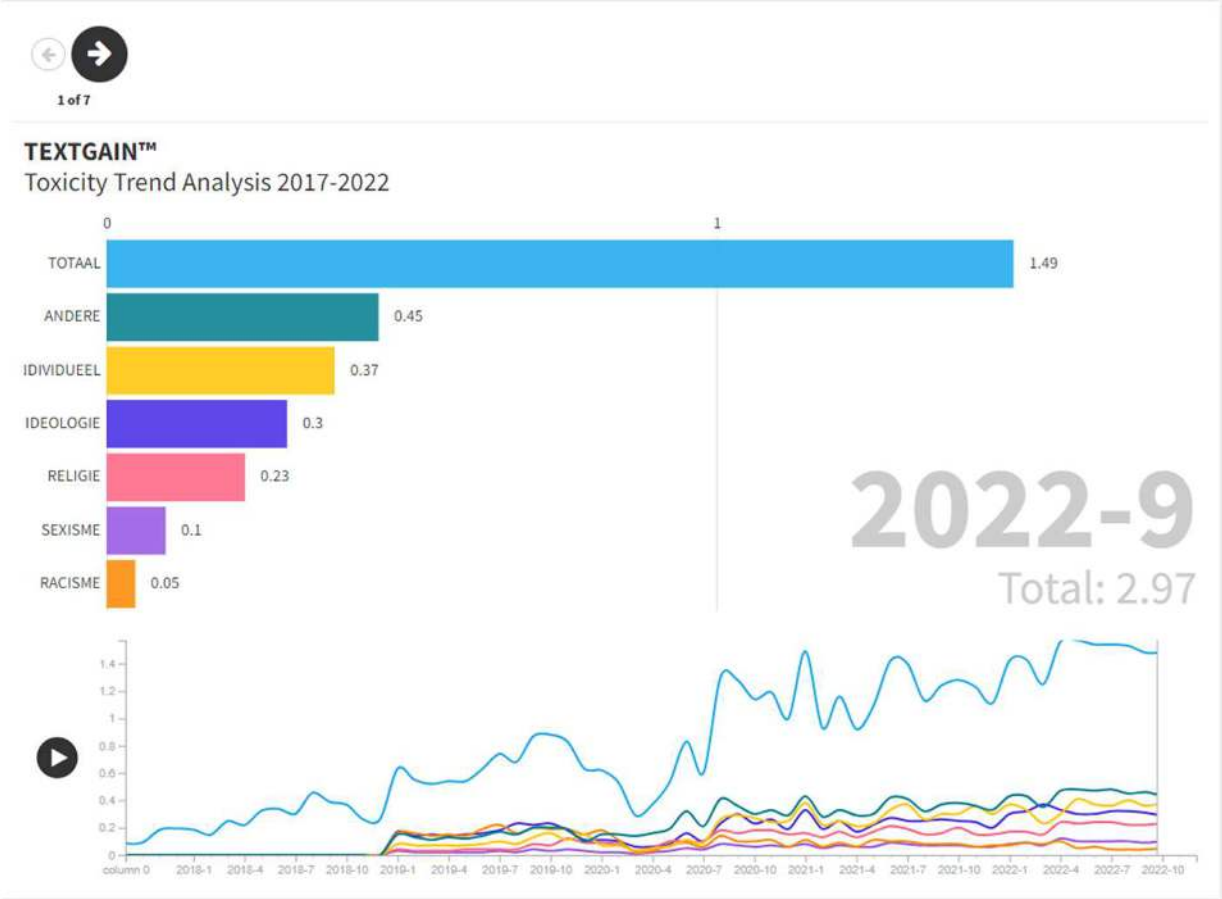
POLARISATIE

Online zingt het vogeltje ranziger dan ooit tevoren

Racistische taal: maal drie. Seksistische taal: maal twee. Dreigende taal: maal drie. Op sociale media gaat het er vandaag veel ranziger aan toe dan vijf jaar geleden.

Stijn Cools
woensdag 17 juni 2020 om 0.00 uur

Trends door de tijd heen



https://public.flourish.studio/story/2112946/

Antwerps AI-bedrijf leidt Europees onderzoek naar online haatspraak

EOOH
EUROPEAN OBSERVATORY OF ONLINE HATE

De bestorming van het Amerikaanse Capitool was een voorbeeld van waar een hardnekkige desinformatiecampagne toe kan leiden. ©AFP

BEN SERRURE | 20 januari 2021 02:55

Meest gelezen

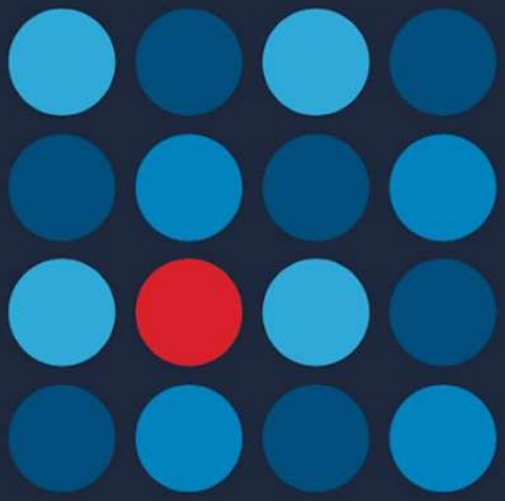
1 Is het stilaan tijd om het feestje op de beurs te verlaten?

2 Vandenbroucke drukt hoop op snelle versoepeling de kop in

3 Heeft Galapagos nog een toekomst?

4 De man die het levenswerk van Leopold Lippens moet voortzetten

5 Jos Sluys, multimiljonair op buikgevoel



EOOH

EUROPEAN OBSERVATORY OF ONLINE HATE

www.eooh.eu

DG JUST • REC-RRAC-AG-2020 • PANORAMA 963801

- **European Observatory of Online Hate : www.EOOH.eu**
- Vroegtijdige detectie van haatzaaiende uitlatingen/desinformatie voor alle 24 Europese talen (+ Arabisch, Russisch, Turks en meer)
- Momenteel monitoren ze **12 social media platforms** via een op maat gemaakt dashboard
- EOOH betrekt meer dan **100 experts bij rondetafelsessies** uit: de academische wereld, het maatschappelijk

Opbouwen van lexicons: Start bij het annotatieproces / categoriseren van toxische taal (legaal en illegaal)

SCORE 0 = no problem, 4 = very problematic

RACISM

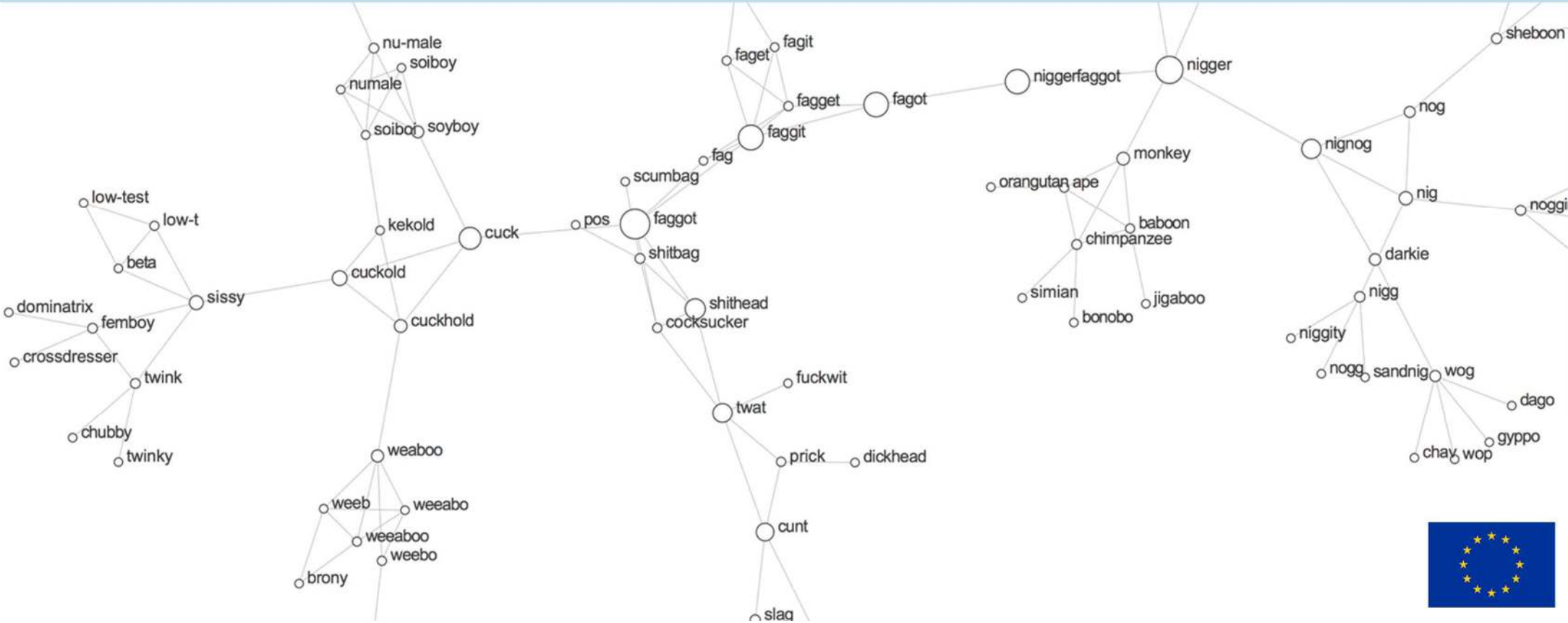
THREATS

1	★	#	WORD			😠	💩	🐸	🤡	👰	🐷	👳🏿	✊🏻	☝🏽	💣
2	0-4	1M	NL		EN	HATE	SHIT	SCUM	FOOL	SLUT	FUCK	GOOK	HEIL	HELL	KILL
9007	1	2	nen blaffer		a gun	☐	☐	☐	☐	☐	☐	☐	☐	☐	☑
9008	2	0	noten op uw zang		talk tough	☑	☐	☐	☐	☐	☐	☐	☐	☐	☑
9009	4	0	stuur ze naar de goelag		send them to the camp	☐	☐	☐	☐	☐	☐	☐	☑	☐	☑
9010	4	0	stuur ze naar de kampen		send them to the camps	☐	☐	☐	☐	☐	☐	☐	☑	☐	☑
9011	1	142	uitgelachen		laughed at	☐	☐	☐	☑	☐	☐	☐	☐	☐	☐
9012	3	0	die zemmertje		that little dick	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
9013	3	0	die zehma		that dick	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
9014	2	24	putain		whore	☐	☑	☐	☐	☐	☐	☐	☐	☐	☐
9015	2	0	denkt dat hij cool		thinks he's cool	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐

thousands of known expressions

universal translation

Machine Learning: automatisch leren / zoeken van nieuwe onbekende uitingen door AI te gebruiken die uitlegbaar is



Beschikbare Social Media platformen die we monitoren

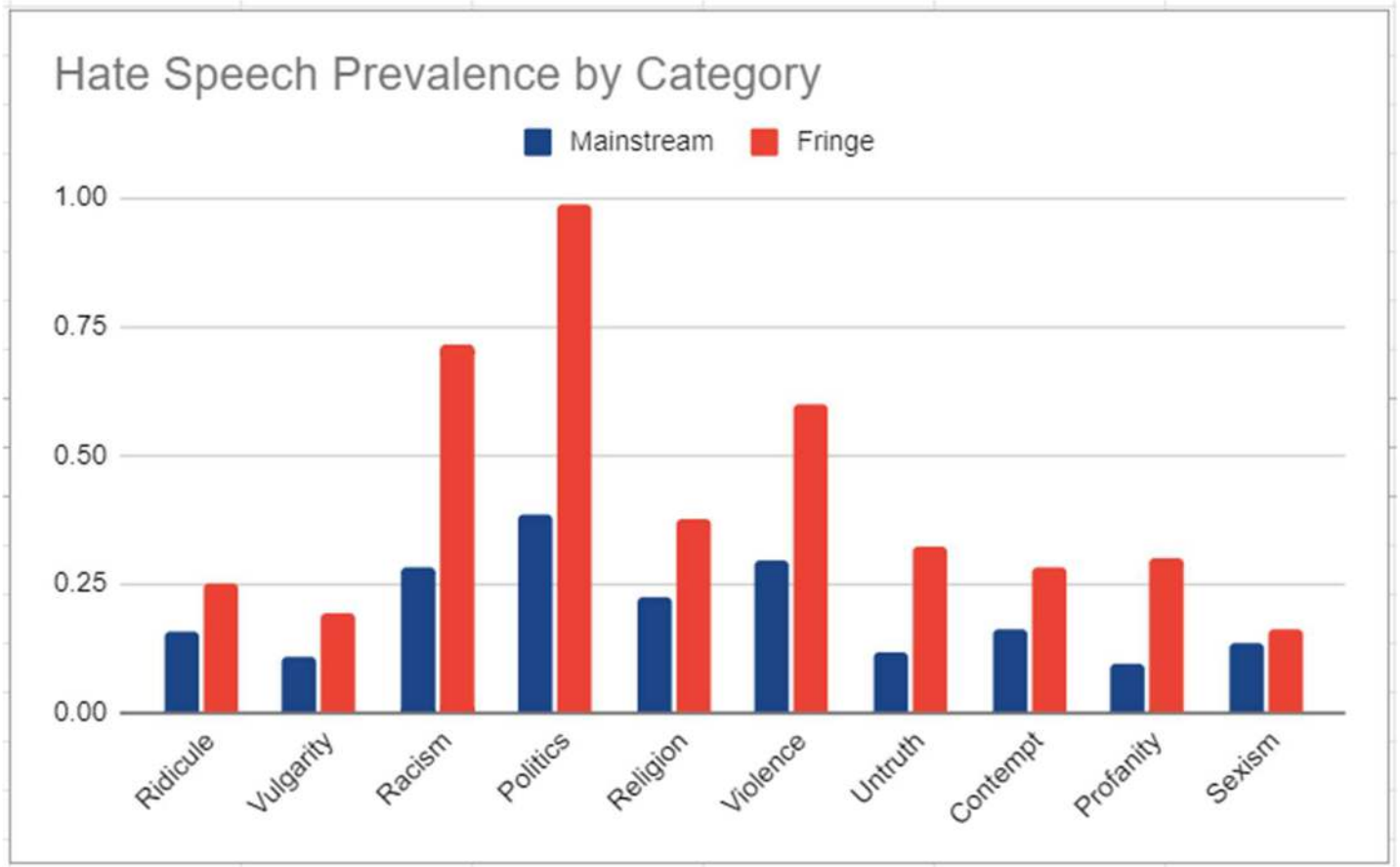
Beschikbaar:

- **X (Twitter)**
- **Telegram**
- **Reddit**
- **Minds**
- **Gab**
- **YouTube**
- **Wordpress sites/blogs**
- **Fourplebs**
- **Facebook**
- **Instagram**
- **TikTok**
- **Google News**

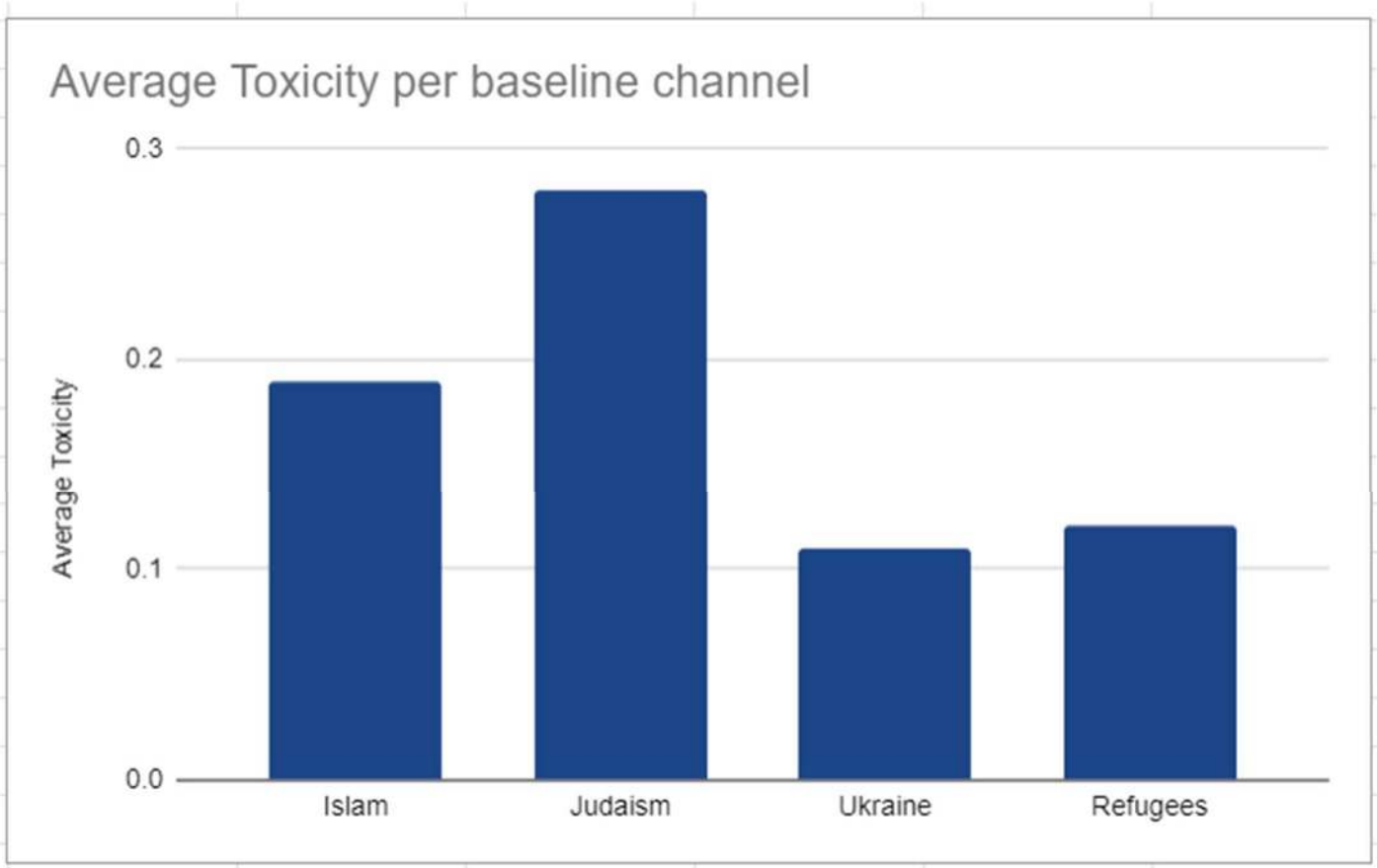
On hold:

- 9gag
- RT (EU ban)
- Vkontakte (EU ban)
- Blogger
- 8kun (offline)

Mainstream vs. Fringe - Hate Speech Prevalence by Category



Baseline channel




deDUIDER

Doelstellingen

Met deDuider stelt Textgain de lokale besturen in staat:

- **Een overzicht** te krijgen van wat er **gaande is rond extremistische uitingen online**.
- Het bieden van diverse **communicatie invalshoeken** om hierop te reageren.
- Het dient als online **vroegdetectie,** signalen detecteren rond extremistische nieuwe trends, thema's en narratieven die zorgwekkend zijn.

We bieden

- **deDuider tool** ter gebruik voor en door het lokaal bestuur.
- **Website** (www.deduider.be) openbaar, met o.a. inzichten en factsheets over nieuwe online fenomenen, plus inlog.
- Een **proactieve communicatiestrategie** (o.a. gericht op de lokale besturen) en via professionele sociale netwerken (zoals LinkedIn) verspreid.
- Een **crisis-incidententeam** (voor urgente vragen).
- Ter info: deDuider tool is nu in de **ontwikkelfase**. Lancering rond eind april 2024.

Waar hebben we AI nodig bij deDuider?

De basis:

- NLP / POW scoren en opbouwen (**Lexicons**) van extreme thema's.
- **Machine learning** (leren vinden van nieuwe trends).

Binnen deDuider:

- **Clusteren** van Social Media data op basis van thema's en invalshoeken.
- **Samenvatten** (summarizing) van comments (LLM).

Opbouwen van Lexicons: Start bij het annotatieproces / categoriseren van toxische taal (legaal en illegaal)

SCORE 0 = no problem, 4 = very problematic

RACISM

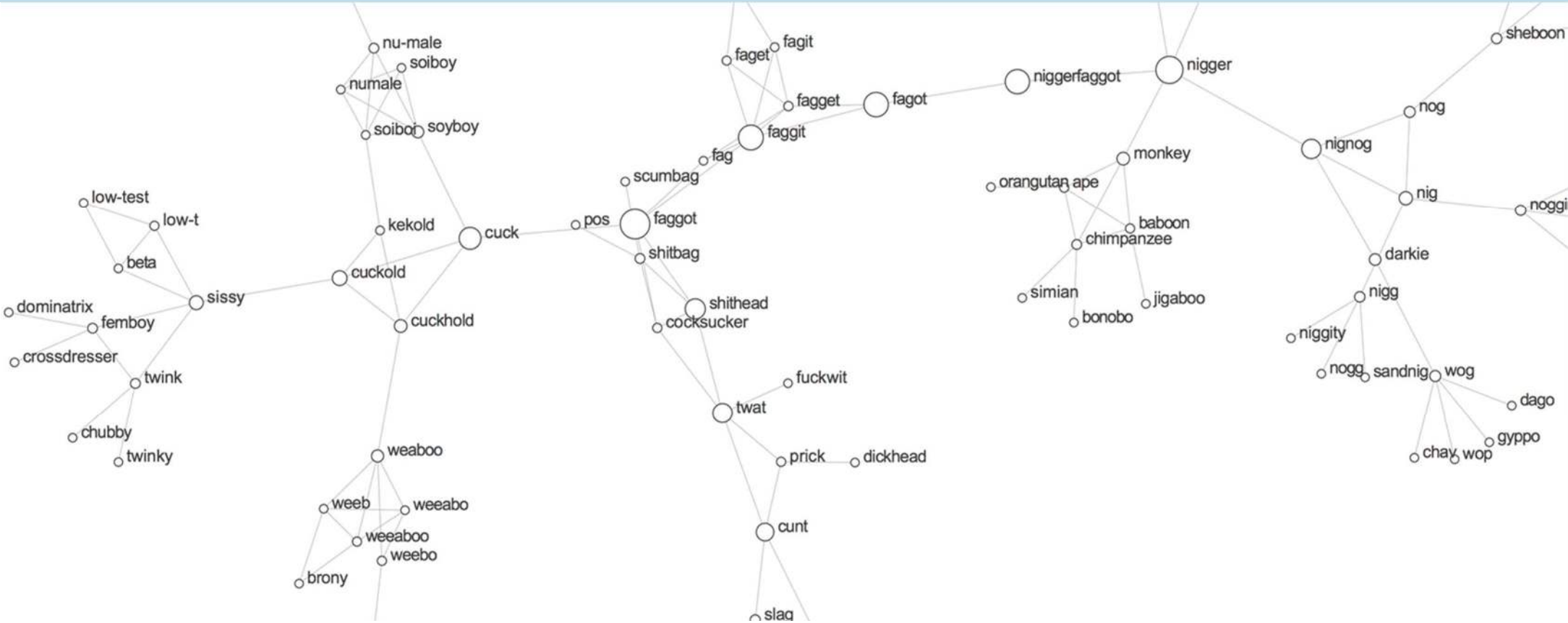
THREATS

1	★	#	WORD												
2	0-4	1M	NL		EN	HATE	SHIT	SCUM	FOOL	SLUT	FUCK	GOOK	HEIL	HELL	KILL
9007	1	2	nen blaffer		a gun	☐	☐	☐	☐	☐	☐	☐	☐	☐	☑
9008	2	0	noten op uw zang		talk tough	☑	☐	☐	☐	☐	☐	☐	☐	☐	☑
9009	4	0	stuur ze naar de goelag		send them to the camp	☐	☐	☐	☐	☐	☐	☐	☑	☐	☑
9010	4	0	stuur ze naar de kampen		send them to the camps	☐	☐	☐	☐	☐	☐	☐	☑	☐	☑
9011	1	142	uitgelachen		laughed at	☐	☐	☐	☑	☐	☐	☐	☐	☐	☐
9012	3	0	die zemmertje		that little dick	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
9013	3	0	die zehma		that dick	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
9014	2	24	putain		whore	☐	☑	☐	☐	☐	☐	☐	☐	☐	☐
9015	2	0	denkt dat hij cool		thinks he's cool	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐

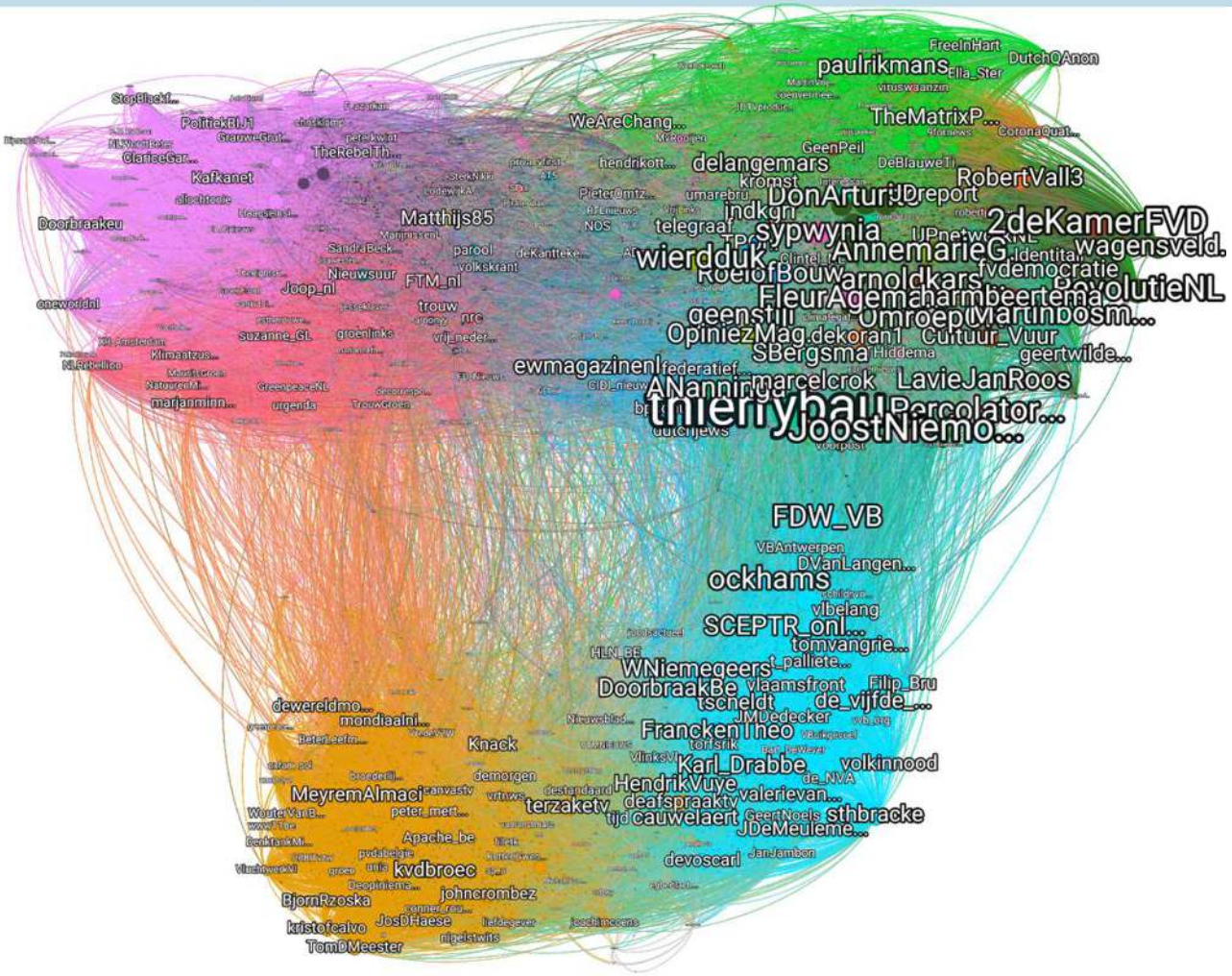
thousands of known expressions

universal translation

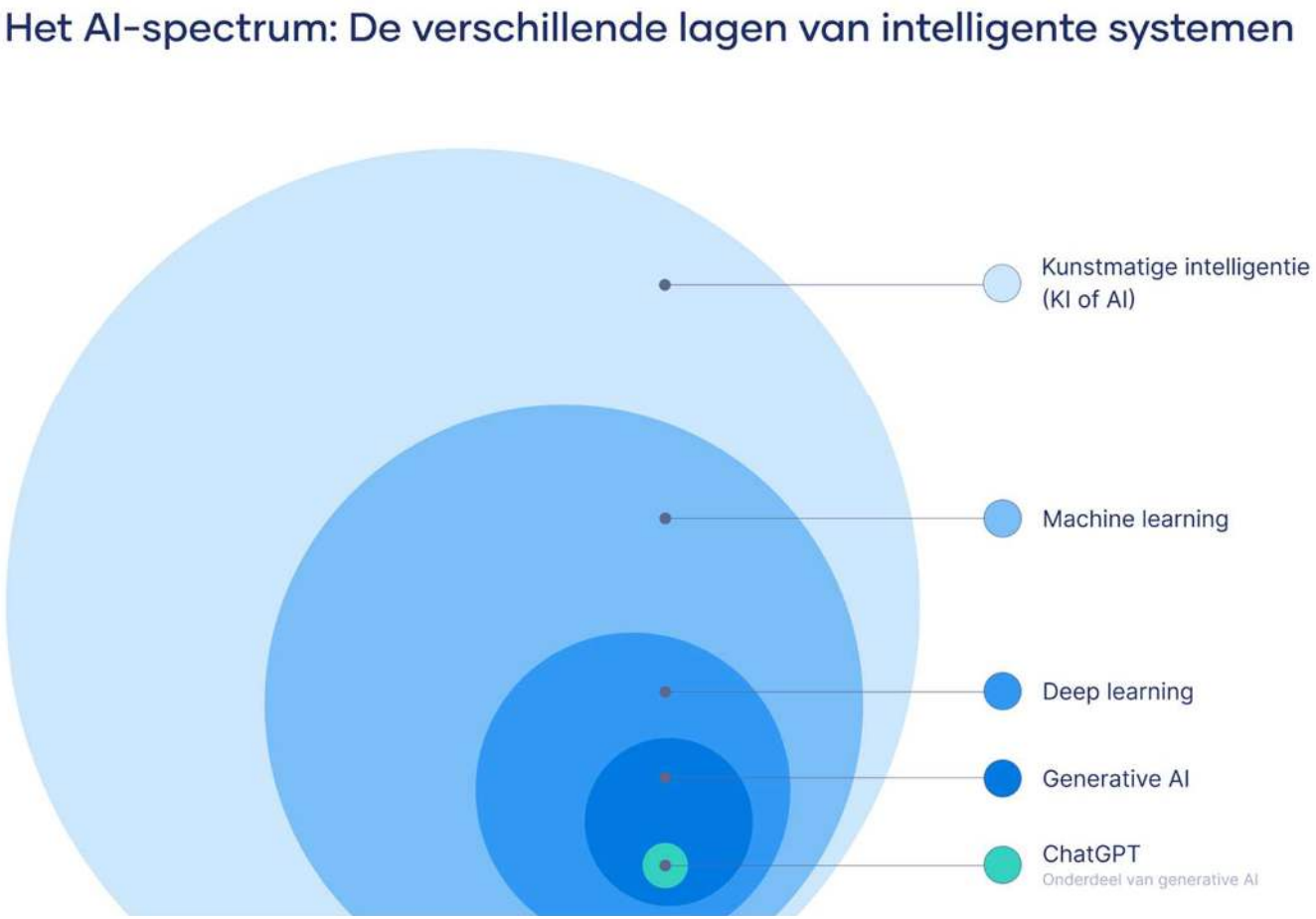
Machine Learning: automatisch leren / zoeken van nieuwe onbekende uitingen door AI te gebruiken die uitlegbaar is



Creatie van clustering:



Samenvatten met modellen zoals ChatGPT:



Risico’s van Large Language Models (LLM)

- **Hallucinatie**, AI kan zaken verzinnen die niet kloppen.
- **Data die de EU uitgaat**, we werken met modellen die op servers in de EU draaien.
- **Black box**, zoveel parameters, dat de uitkomst moeilijk te achterhalen is.
- **(in de uitvoering) Overspoeld worden** met nepnieuws en dog whistles.

Risico’s van Large Language Models (LLM)

Wat zijn de risico’s:

- **Hallucinatie**
- > Human in the loop
- > Fine tuning van output

Risico’s van Large Language Models (LLM)

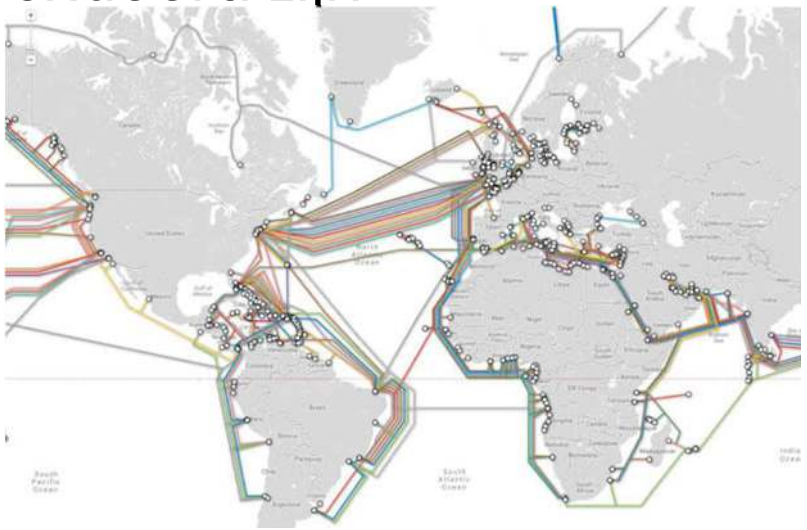
Wat zijn de risico’s:

- Hallucinatie
- **> Human in the loop**
- **> Fine tuning van output**

Risico’s van Large Language Models (LLM)

Wat zijn de risico’s:

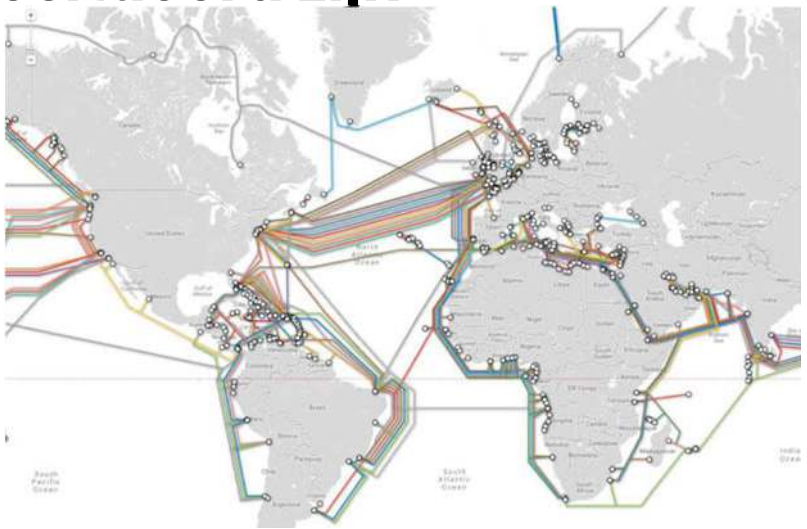
- **Data die de EU uitgaat**
- > Alleen werken met LLM die in de EU gesitueerd zijn
- > On-premises AI-modellen draaien



Risico’s van Large Language Models (LLM)

Wat zijn de risico’s:

- Data die de EU uitgaat
- **> Alleen werken met LLM die in de EU gesitueerd zijn**
- **> On-premises AI-modellen draaien**



Risico’s van Large Language Models (LLM)

Wat zijn de risico’s:

- **Black box van AI?**
- > AI-modellen (door)ontwikkelen die de uitkomsten eXplainable kunnen maken (xAI)

Risico’s van Large Language Models (LLM)

Wat zijn de risico’s:

- Black box van AI?
- **> AI-modellen (door)ontwikkelen die de uitkomsten eXplainable kunnen maken (xAI)**

Risico’s van Large Language Models (LLM)

Wat zijn de risico’s:

- **Overspoeld worden met o.a. Nepnieuws en dog whistles**
- > Detectie doorontwikkelen
- > Educatie opvoeren

Risico’s van Large Language Models (LLM)

Wat zijn de risico’s:

- Overspoeld worden met o.a. Nepnieuws en dog whistles
- **› Detectie doorontwikkelen**
- **› Educatie opvoeren**

SAVE the DATE

Safer Internet Day

2024 | Tuesday 6 February

www.saferinternetday.org

European Commission

INHOPE

ins@fe

Textgain Academy - Safer Internet Day 2024 - 6 februari

- **Programma 6 februari:**
- 09.00 uur: **Workshop OSINT**, meer dan gewoon wat googlen
- 10.00 uur: De opkomst van de **soevereine beweging** (en autonome burgers)
- 12.00 uur: **Hatemoji**

- Voor inschrijvingen en meer info, kijk op:

deDUIDER

www.deduider.be

Interesse: laat je e-mail alvast achter

Lancering rond eind april 2024

deDUIDER

Gijs van Beek

deDuider (www.deduider.be)

Textgain (www.textgain.com)

Email: Gijs@textgain.com

Mobile: +316 273 422 68

textgain

Artificial Intelligence that reads between the lines

Thank you!