De Duider

Guy De Pauw, Olivier Cauberghs

13 November 2023

textgain

Artificial Intelligence that reads between the lines

textgain

- Opgericht in 2015
- Spin-off van de Universiteit Antwerpen
- Artificiële Intelligentie:
  - Zelflerende systemen
  - Natural Language Processing
- *AI for Good*: Social media Monitoring

2016: IS-propaganda

DeMorgen.

POLITIEK | MENINGEN | OORLOG IN OEKRAÏNE | TV & CULTUUR | TECH

strijd tegen IS

Met deze software wil Antwerps bedrijf IS-propaganda opsporen

Guy De Pauw en Tom De Smedt van het Antwerpse taaltechnologiebedrijf Textgain. Beeld Tine Schoemaeker

In de strijd met online IS-propaganda heeft het Antwerpse taaltechnologiebedrijf Textgain een nieuw computerprogramma ontwikkeld dat foto's gepost door IS-aanhangers herkent. "We willen meehelpen om radicalisering sneller in kaart te brengen."

RANI DECOCK 8 september 2016, 20:49

JIHAD	holy war	79%
KAFIR, KUFFAR, KUFR	unbeliever(s)	98%
SHIRK	idolatry = worshipping other gods but Allah	96%
MUHAJIR	jihad warrior from a foreign country	80%
LION	jihad warrior with bravery	81%
MURTAD	a bad Muslim raised by good Muslims	95%
TAKFIR	a bad Muslim	86%
TAKBIR	*response* = "Allahu Akbar!", used in great determination	88%
SHARIA	*roughly*: religious law	88%
UMMA	*roughly*: the unity of all Muslims across states	90%
DAWLA	*roughly*: a temporary state to promote Umma	66%
COCONUT	takfir (i.e., brown outside but white inside)	86%
DOGS, PIGS	US or European Christians or atheists	75%

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

AMiCA

textgain


DeMorgen.
strijd tegen IS
Met deze software wil Antwerps bedrijf IS-propaganda opsporen
Guy De Pauw en Tom De Smedt van het Antwerpse taaltechnologiebedrijf Textgain. Beeld Tine Schoemaeker

PROJECT GREY

factcheck. vlaanderen

DTCT
DETECT THEN ACT

RHETORiC

IMSyPP

COMMIT
COMMunication campaign against extremism and radicalisation

EOOH
EUROPEAN OBSERVATORY OF ONLINE HATE

textgain academy

BENEDMO

INSIGHT

EHBT

RHE:PLUS

1995 2012 2015 2016 2017 2018 2019 2020 2021 2022

PROJECT GREY

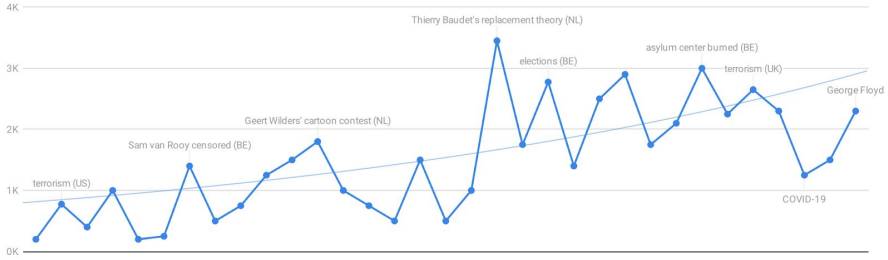

De Standaard Meest recent Binnenland Coronacrisis Meer ▾

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.

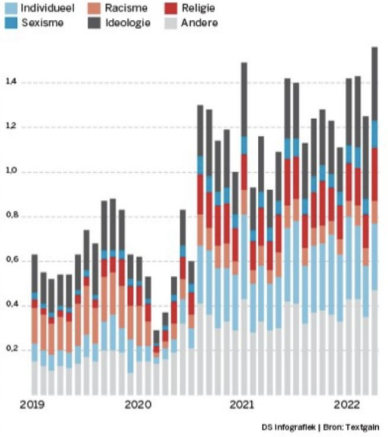
Stijn Cools
woensdag 17 juni 2020 om 0.00 uur



Op Twitter heerste even vrede. Lang heeft dat niet geduurd

HAATSPRAAK Het aandeel giftige berichten, persoonlijke verwijten en discriminerende uitlatingen heeft nieuwe hoogtes bereikt op socialenetwerksite Twitter.

Aantal toxische berichten online
Januari 2019 - april 2022, in procent



DS Infografiek | Bron: Textgain

De uitbraak van de coronacrisis, in maart twee jaar geleden, had een verrassend effect op de toxiciteit van socialemediasite Twitter. Het aandeel toxische berichten ging prompt naar beneden. Enkele maanden later klommen de cijfers evenwel met gemak tot boven het niveau van voor de uitbraak en zelfs tot nieuwe hoogtes. Het gevoel van samenhorigheid was snel weer voorbij.

Tussen januari 2019 en april 2022 heeft het Belgische technologiebedrijf Textgain, een spin-off van de UAntwerpen, maandelijks één tot anderhalf miljoen Nederlandstalige berichten gescreend op hun toxiciteit. Dat gebeurde samen met enkele Europese universiteiten in het kader van een door de Europese Commissie gefinancierd project om haatspraak te monitoren, te voorkomen en te bestrijden, genaamd IMSyPP.

Met de hulp van kunstmatige intelligentie legde Textgain een databank met ongeveer 12.500 toxische termen aan. Die werden gebruikt om patronen van woorden in berichten op Twitter te labelen als al dan niet toxisch.

Polariserende inhoud

De termen uit de databank variëren van ontmenselijkend tot discriminerend of polariserend. Het gaat bijvoorbeeld om een racistische benaming als 'dobberneger', een synoniem voor mensen die per boot hun land ontvlucht zijn. Antisemitische uitlatingen zoals 'idijood' of 'haakneus' maken eveneens deel uit van de databank, naast het meer gebruikelijke 'smeerlap'. Ook verwijten aan het Vlaamsgezinde gedeelte van het politieke spectrum ontbreken niet, zoals 'Vlazi' of 'goedvlaem'.

'Het is niet dat iedere tweet met bijvoorbeeld de term "smeerlap" automatisch als toxisch gelabeld wordt. We filteren op bepaalde patronen van tekst om de meest toxische eruit te halen. Als bijvoorbeeld een woord tussen aanhalingstekens staat en dus geciteerd wordt, dan weet ons algoritme dat daar geen rekening mee gehouden moet worden', zegt Guy De Pauw, de ceo van Textgain. 'Behalve als de aanhalingstekens wel kunnen wijzen op toxiciteit, zoals wanneer het om "jongeren" gaat.'

Het resultaat van die screening toont dat, ondanks de inspanningen vanuit overheden, middenveldorganisaties en Twitter zelf, de toxiciteit blijft toenemen. April 2022 bleek bijvoorbeeld de meest toxische maand te zijn sinds het begin van het onderzoek. Meer dan 1,5 procent van de onderzochte berichten bevat inhoud die volgens de databank van Textgain polariserend is. Dat is meer dan een verdubbeling tegenover de eerste maand die deel uitmaakt van de analyse, namelijk januari 2019.

Tegelijk klinkt 1,5 procent weinig – en valt het dus nog wel mee met de giftigheid op Twitter. Daar wil De Pauw geen uitspraken over doen. Hij legt uit dat het algoritme de lat voor het label toxisch hoog legt om valse positieven te vermijden.

Door de zoektermen onder te brengen in categorieën, laat het algoritme van Textgain zien waar de problemen zich situeren. In de screening van vorige maand is de restcategorie de grootste. 'In deze categorie zijn alle toxische berichten over corona ondergebracht, net zoals kritiek op de media', legt De Pauw uit.

Knokkerokko

De categorie racisme daarentegen is kleiner dan men zou verwachten. De Pauw: 'Maar dat beeld is misschien wat vertekend omdat bijvoorbeeld islamofobie in de categorie religie is ondergebracht.'

De toxiciteit online staat niet los van de actualiteit, legt De Pauw uit. De grote sprong tussen juli en augustus 2020 in de grafiek verklaart hij onder meer door de rellen op het strand van Blankenberge en Knokke –op sociale media omgedoopt tot 'Knokkerokko' –, rellen in de Schilderswijk in Den Haag en de bijkomende maatregelen die de Antwerpse provinciegouverneur Cathy Berx nam om de verspreiding van het coronavirus in te perken.

> Ondanks de inspanningen vanuit overheden, middenveldorganisaties en Twitter zelf, blijft de toxiciteit op Twitter toenemen

De voorbije jaren kwamen eerdere onderzoeken van Textgain al ter sprake in de commissie Radicalisering van het Vlaams Parlement en de Senaat. Toen werd vastgesteld dat de Vlaamse sociale media niet zozeer gebukt gaan onder georganiseerd extremisme, als wel onder tal van gepolariseerde persoonlijke meningen, 'waarbij sommige burgers dreigende taal niet schuwen'.

Kunstmatige intelligentie

Met deze studie heeft Textgain niet alleen geobserveerd wat er leeft op Twitter, maar heeft het bedrijf ook verder geïnnoveerd met taaltechnologie en kunstmatige intelligentie. De komende jaren leidt Textgain voor de Europese Commissie het European Observatory of Online Hate, dat de dynamiek van online haat beter in kaart moet brengen, werkt het samen met vzw's zoals Dare to be grey om die online haat te counteren en onderzoekt het voor Facebook de populariteit van bepaalde complottheorieën.

Stijn Cools

IMSyPP

TCT

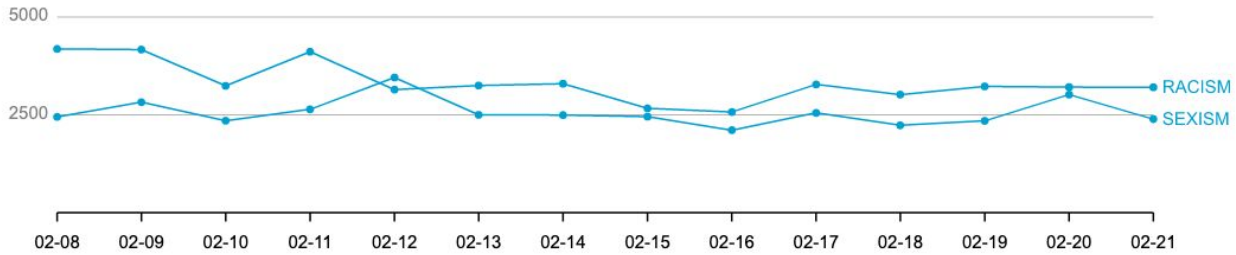
TROLLFEED
Feed the trolls.

EN DE FR NL HU

Sort by: score

Look for: edit #Antifa

ALL	RACISM	SEXISM	THREAT	REPLIES	REPORTS
3.71M	1.1M	524K	317K	310	295



PAST HOURS

bitch

stfu

skank

NEWS Sat Feb 20 21:40:16 2021 Kangaroo Island dunnart: Saving a bushfire-ravaged marsupial read

TYRONE-PLEADS-WITH-HIS-MAD-LAWYER

Mon Feb 22 01:45:26 2021

PROFANITY RIDICULE DEHUMANIZATION

you weird as fuck this whole shit was dropped and now you back calling me a random bozo go eat a fat cock you piece of shit http://...

REACH	SCORE	RACISM	SEXISM	THREAT
●●○○	●●●●	●○○○	○○○○	○○○○

reply report ×

ANONYMIZED

Sun Feb 21 09:28:31 2021

TCT

TROLLFEED
Feed the trolls.

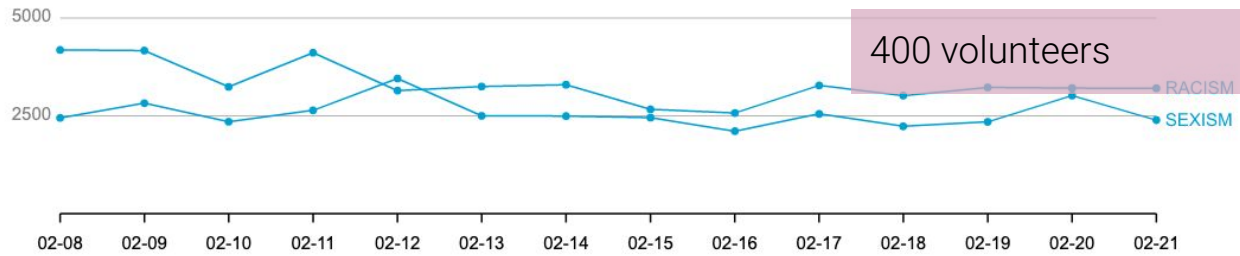
EN DE FR NL HU

Sort by: score

Look for: edit #Antifa

MEMES

ALL	RACISM	SEXISM	THREAT	REPLIES	REPORTS
3.71M	1.1M	524K	317K	310	295



400 volunteers

PAST HOURS

bitch

stfu

skank

NEWS Sat Feb 20 21:40:16 2021 Kangaroo Island dunnart: Saving a bushfire-ravaged marsupial read

TYRONE-PLEADS-WITH-HIS-MAD-LAWYER

Mon Feb 22 01:45:26 2021

PROFANITY RIDICULE DEHUMANIZATION

you weird as fuck this whole shit was dropped and now you back calling me a random bozo go eat a fat cock you piece of shit http://...

REACH	SCORE	RACISM	SEXISM	THREAT
●●○○	●●●●	●○○○	○○○○	○○○○

reply report ×

ANONYMIZED

Sun Feb 21 09:28:31 2021

TCT

TROLLFEED
Feed the trolls.

EN DE FR NL HU

Sort by:
score

Look for: edit
#Antifa

MEMES

ALL
3.71M

THREAT
317K

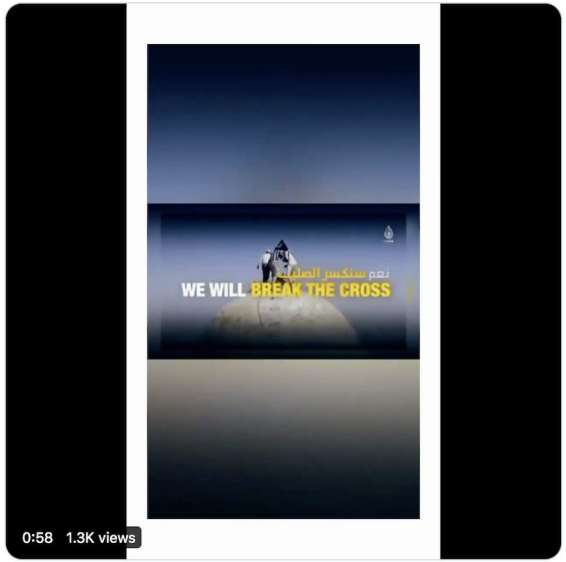
REPLIES
310

REPORTS
295

Top Latest People Photos Videos

O≣sama @udujfud · 32m
O infidels of the world

#Vienna



4

5000
2500
02-08
02-17
02-18
02-19
02-20
02-21
RACISM
SEXISM

PAST HOURS
bitch
stfu
skank

NEWS Sat Fe ravaged marsupial read

TYRON
Mon
PROFA
you w
callin

REACH SCORE RACISM SEXISM THREAT
●●○○ ●●●● ●○○○ ○○○○ ○○○○

reply report ×

ANONYMIZED
Sun Feb 21 09:28:31 2021

TCT

TROLLFEED
Feed the trolls.

EN DE FR NL HU

Sort by:
score

Look for: edit
#Antifa

MEMES

ALL
3.71M

5000
2500
02-08 02-

NEWS Sat Fe

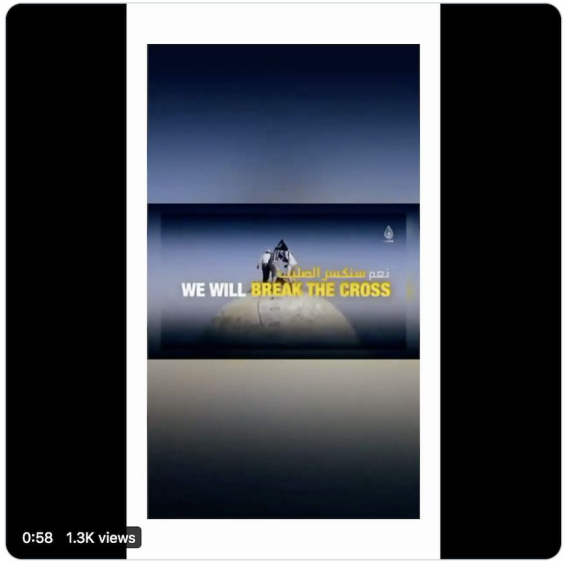
TYRON
Mon
PROFA
you w
callin

ANONYMIZED
Sun Feb 21 09:28:31 2021

Top Latest People Photos Videos

O≣sama @udujfud · 32m
O infidels of the world

#Vienna



4

Your report

An update on your reports

Thanks again for letting us know. Our investigation found these accounts violated the Twitter Rules:

Mr.TweeTy
@tw33tz0r

- Violating our rules against posting media depicting the moment of death of an individual(s).

Osama
@sss2222i90

- Violating our rules against posting media depicting gratuitous gore.


Waar beleggers deze week naar uitkijken


Parket start onderzoek naar Dries Van Langenhove


Vijf favoriete aandelen van Rob Siegersma


Ook in lockdown kiezen almaar meer bedienden voor bedrijfswagen


Corona doet psychol wachtlijsten afsluiter

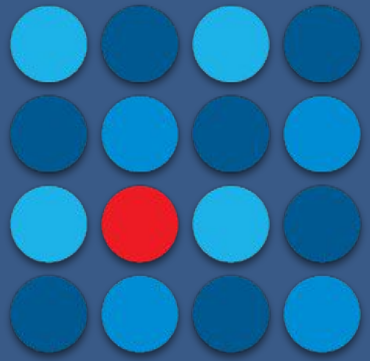
Antwerps AI-bedrijf leidt Europees onderzoek naar online haatspraak

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

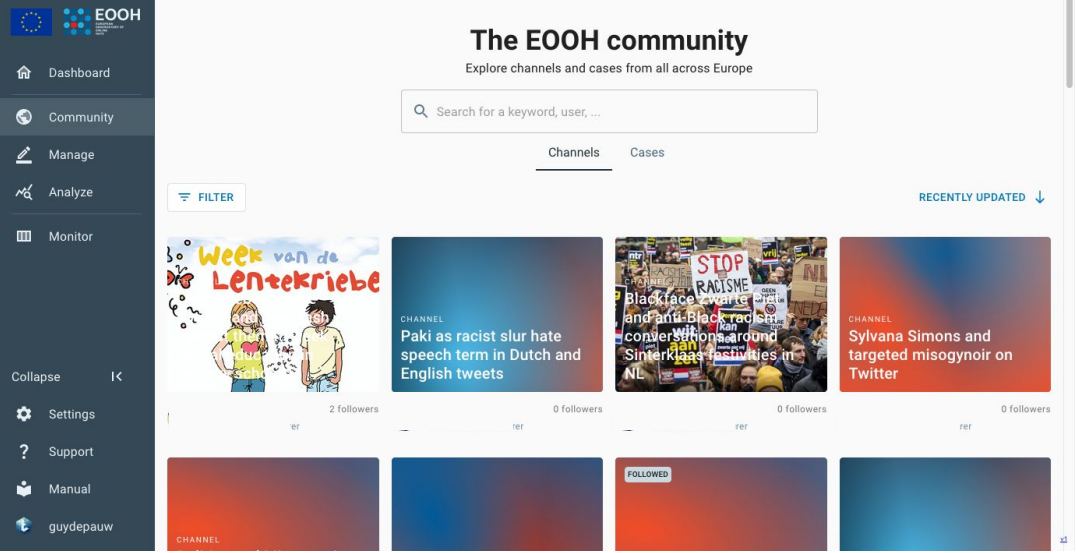
1 Is het stilaan tijd om het feestje op de beurs te verlaten?
2 Vandenbroucke drukt hoop op snelle versoepeling de kop in
3 Heeft Galapagos nog een toekomst?
4 De man die het levenswerk van Leopold Lippens moet voortzetten
5 Jos Sluys, multimiljonair op buikgevoel

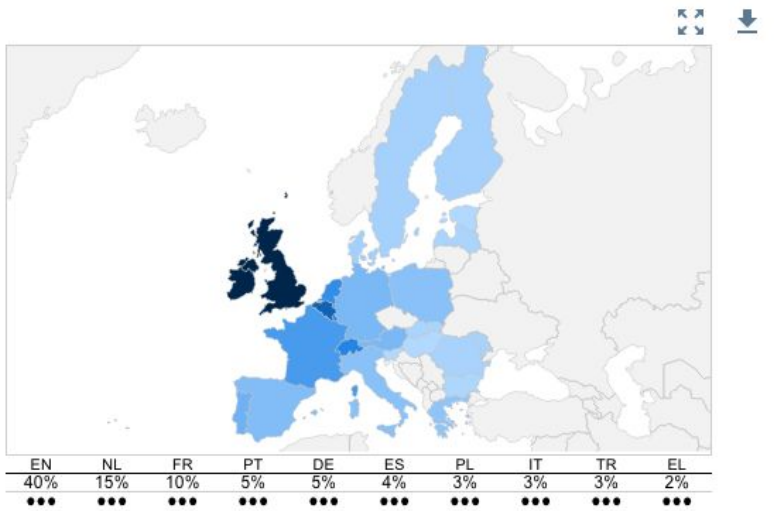


EOOH

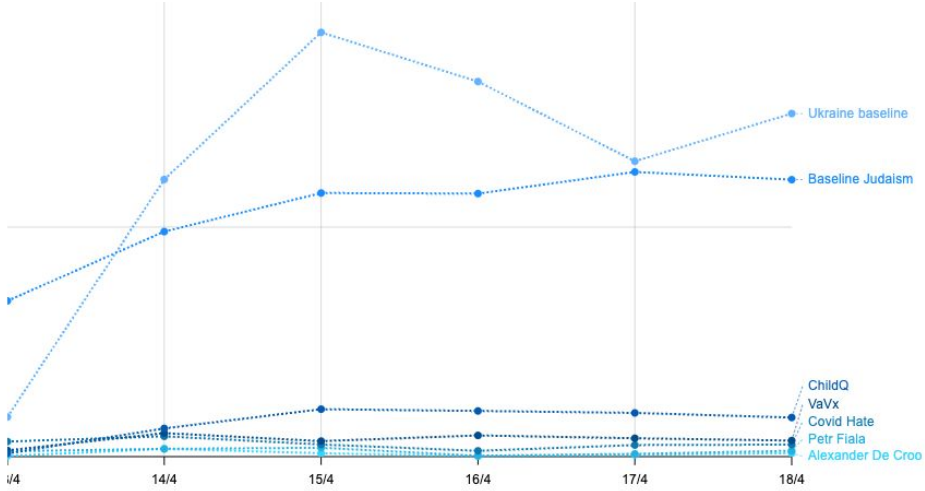
EUROPEAN OBSERVATORY OF ONLINE HATE

- Haatspraak detectie voor alle 24 talen van de EU
- 50 experten in 13 rondetafelsessies
  - Civil Society (Hanna Arendt, ...)
  - Law inforcement (CUTA, ...)
  - Social Media (Twitter, TikTok, ...)

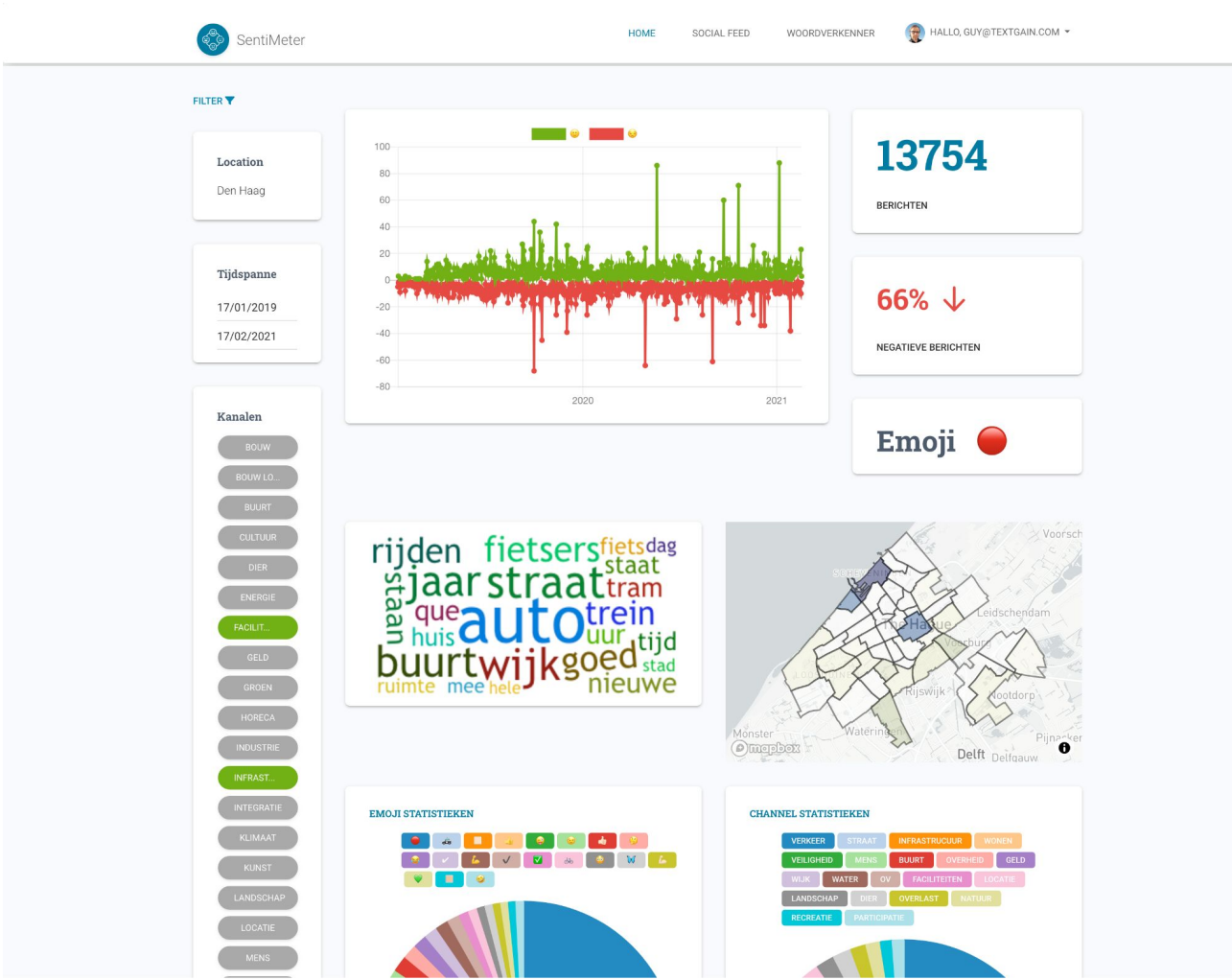




	category	#	score	keywords
🤡	RIDICULE	●●●	0.55	*judíos, türk, covidiot*
🐸	CONTEMPT	●●●	0.50	*nazi, ser, vers*
🙈	RACISM	●●●	0.40	*jews, juif, juifs*
🤦‍♀️	SEXISM	●●●	0.55	*tant, ho, violence*
✊	POLITICS	●●●	0.35	*nazi, nazis, zemmour*
☝️	RELIGION	●●●	0.40	*jews, juif, juifs*
💣	THREAT	●●●	0.30	*rape, guerra, genocide*
👽	UNTRUTH	●●●	0.40	*zionist, nom, żydzi*
✡️	JEWS	●●●	0.40	*nazi, antisemitism, hitler*
✝️	CHRISTIANS	○○○	0.00	
☪️	MUSLIMS	○○○	0.00	



sentimeter


SentiMeter
HOME
SOCIAL FEED
WOORDVERKENNER
HALLO, GUY@TEXTGAIN.COM
FILTER
Location
Den Haag
Tijdspanne
17/01/2019
17/02/2021
Kanalen
BOUW
BOUW LO...
BUURT
CULTUUR
DIER
ENERGIE
FACILIT...
GELD
GROEN
HORECA
INDUSTRIE
INFRAST...
INTEGRATIE
KLIMAAT
KUNST
LANDSCHAP
LOCATIE
MENS
100
80
60
40
20
0
-20
-40
-60
-80
2020
2021
13754
BERICHTEN
66%
NEGATIEVE BERICHTEN
Emoji
rijden
fietsers
fiets
dag
staat
jaar
straat
tram
staan
que
auto
trein
huis
uur
tijd
buurt
wijk
goed
stad
ruimte
mee
hele
nieuwe
Voorsch
Leidschendam
Rijswijk
Pijnacker
Delft
Delfgauw
mapbox
EMOJI STATISTIEKEN
CHANNEL STATISTIEKEN
VERKEER
STRAAT
INFRASTRUCUUR
WONEN
VEILIGHEID
MENS
BUURT
OVERHEID
GELD
WIJK
WATER
OV
FACILITEITEN
LOCATIE
LANDSCHAP
DIER
OVERLAST
NATUUR
RECREATIE
PARTICIPATIE

deDuider.be

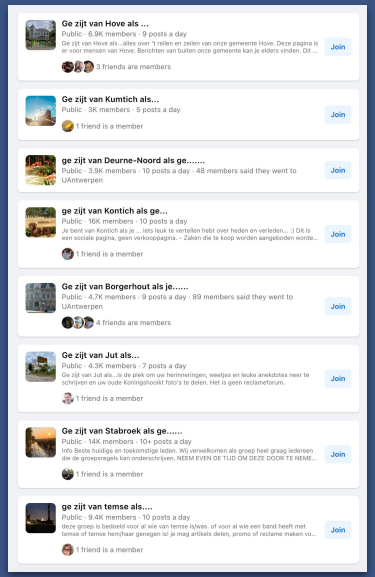
De Duider

- Sociale Mediastromen zijn te groot om zelf te volgen
- Online en offline wereld vinden elkaar steeds meer
- Vragen:
  - Hoe als lokaal bestuur het overzicht bewaren?
  - Trends vs anecdotes
  - Hoe (nieuwe) zorgwekkende maatschappelijke fenomenen in kaart brengen?
    - (online) haatspraak
    - desinformatie

De Duider: (hyperlokale) social media monitoring van zorgwekkende fenomenen zonder de privacy van de burgers te schenden

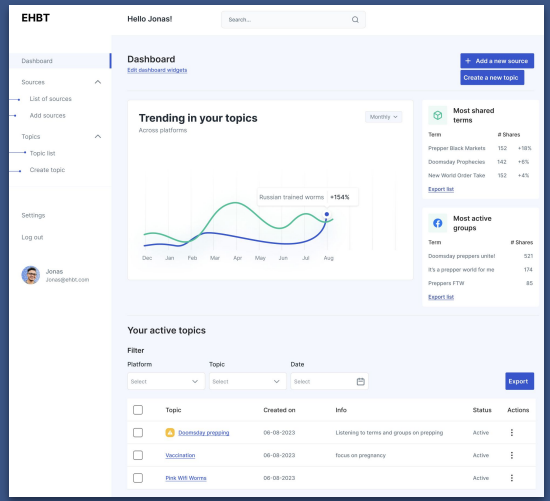
De Duider

- Social-media monitoring
  - Regionaal niveau: reacties op nieuwsmedia, telegram-groepen, …
  - Monitoren op hyperlokaal niveau



De Duider

- Social-media monitoring
- Dashboard
  - Overzicht van regionale en lokale trends
  - Samenvatting van meningen rond zorgwekkende thema's



Gelijkheid tussen man en vrouw

Seksualisering van vrouwenlichamen (55)

Uit onderzoek blijkt dat blote borsten geen choquerend effect hebben. Borsten zijn immers een natuurlijk onderdeel van het voedingsproces van kindjes. Het is tijd voor minder preutsheid en meer gendergelijkheid. Vrouwen moeten zelf kunnen beslissen over hun lichaam en de seksualisering van vrouwenlichamen moet stoppen.

Seksualisering van vrouwenlichamen

Vrijheid van individuele keuze

Sociale normen en fatsoen

Persoonlijke vrijheid en individualisme

De Duider

- Social-media monitoring
- Dashboard
  - Login
  - Overzicht van regionale en lokale trends
  - Samenvatting van meningen rond thema’s

cocreatie

- Informatie- en advies:
  - Duiding bij (nieuwe) fenomen door Textgain Academy
  - Crisisincidententeam

De Duider: tijdlijn

- Samenstelling klankbordgroep **2023-Q4**
  - Communicatiemedewerkers
  - Subject matter experts
  -
- Ideation Sessions met KBG **2024-Q1**
  - Scope en bronnen
  - Functionaliteit dashboard
- Annotatie van trefwoordenlijst **2024-Q1/Q2**
- Dashboard live **2024-Q2**
  - Iteraties **2024-Q3 → 2025-Q4**
- Informatiedeling **2024-Q3 → 2025-Q4**
  - Tutorials, intake
  - Duiding & updates (incl rapid response)
  - Crisisincidententeam

Interesse?
info@deDuider.be

Het Team

Guy De Pauw
CEO, co-founder

Tom De Smedt
CTO, co-founder

Redouan El Hamouchi
COO, co-founder

Gijs van Beek
CBDO, co-founder

Elizabeth Cappon
data scientist

Pierre Voué
data scientist

Andrew Kosar
data scientist

Lisa De Smedt
data scientist

Lydia El Khouri
Civil Society Outreach

Dora Modrall Sperling
Front-end Developer

Olivier Cauberghs
Textgain Academy

Walter Daelemans
co-founder