Technologie en P/CVE

Olivier Caubergh s

textgain

Artificial Intelligence that reads
between the lines

EOOH

EUROPEAN
OBSERVATORY OF
ONLINE
HATE

EOOH dashboard tech roadmap

Olivier Cauberghs
www.eooh.eu

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

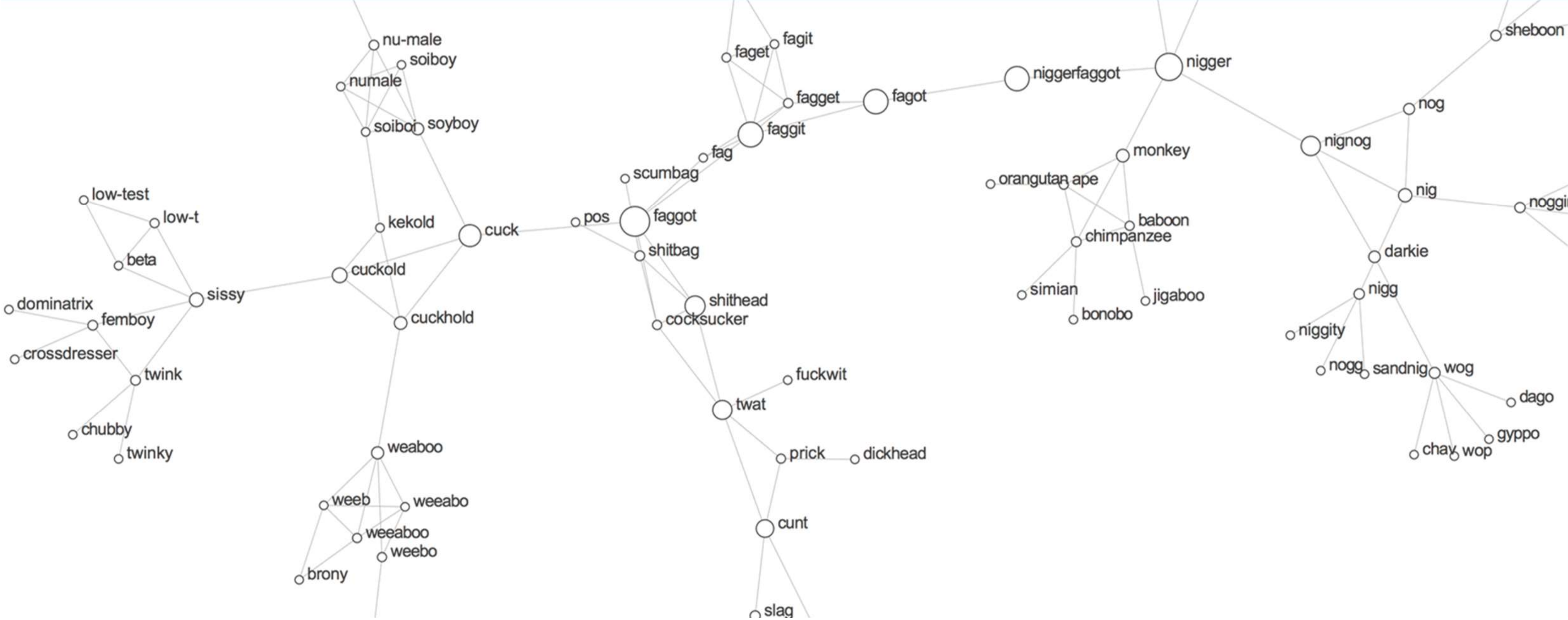
TRANSPARENT AI

- Technology that detects problematic content can help moderators
- Technology that learns by itself (AI) can also be prejudiced however
- That’s why Textgain will develop new transparent AI for EOOH
  - compliant with the EU’s GDPR privacy regulations
  - compliant with the EU’s AI & Ethics recommendations
- For all **24 languages** in the EU + Arabic, Turkish and Russian

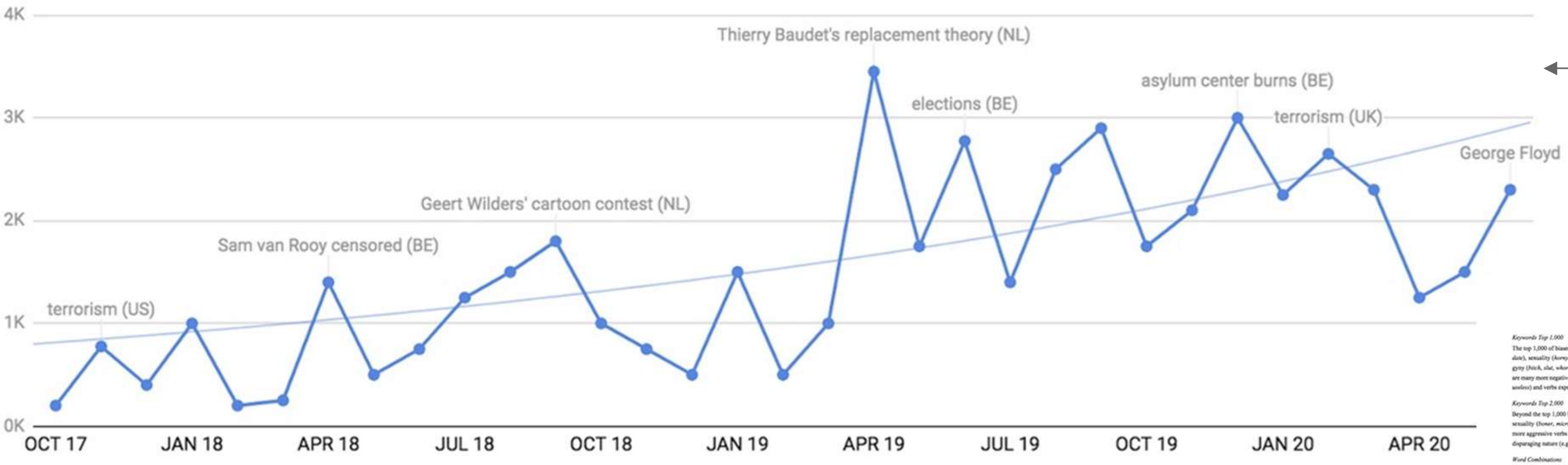
ONLINE TOOL CONNECTS TO PUBLIC APIs (25)

Telegram too

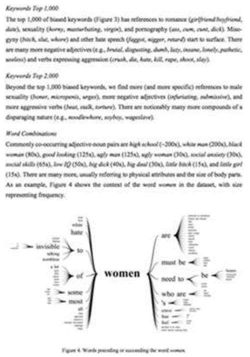
CONTINUOUSLY LEARNS NEW UNKNOWN EXPRESSIONS



ONLINE TOOL GENERATES TREND REPORTS



rise of extremism in the lowlands



Keywords Top 1,000

The top 1,000 of biased keywords (Figure 3) has references to romance (*girlfriend/boyfriend, date*), sexuality (*horny, masturbating, virgin*), and pornography (*ass, cum, cunt, dick*). Misogyny (*bitch, slut, whore*) and other hate speech (*faggot, nigger, retard*) start to surface. There are many more negative adjectives (e.g., *brutal, disgusting, dumb, lazy, insane, lonely, pathetic, useless*) and verbs expressing aggression (*crush, die, hate, kill, rape, shoot, slay*).

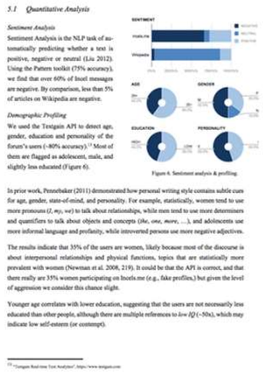
Keywords Top 2,000

Beyond the top 1,000 biased keywords, we find more (and more specific) references to male sexuality (*boner, micropenis, urges*), more negative adjectives (*infuriating, submissive*), and more aggressive verbs (*beat, stalk, torture*). There are noticeably many more compounds of a disparaging nature (e.g., *noodlewhore, soyboy, wageslave*).

Word Combinations

Commonly co-occurring adjective-noun pairs are *high school* (~200x), *white man* (200x), *black woman* (80x), *good looking* (125x), *ugly man* (125x), *ugly woman* (30x), *social anxiety* (30x), *social skills* (65x), *low IQ* (50x), *big dick* (40x), *big deal* (30x), *little bitch* (15x), and *little girl* (15x). There are many more, usually referring to physical attributes and the size of body parts. As an example, Figure 4 shows the context of the word *women* in the dataset, with size representing frequency.

Figure 4. Words preceding or succeeding the word women.



5.1 Quantitative Analysis

Sentiment Analysis

Sentiment Analysis is the NLP task of automatically predicting whether a text is positive, negative or neutral (Liu 2012). Using the Pattern toolkit (75% accuracy), we find that over 60% of Incel messages are negative. By comparison, less than 5% of articles on Wikipedia are negative.

Demographic Profiling

We used the Textgain API to detect age, gender, education and personality of the forum's users (~80% accuracy).[13] Most of them are flagged as adolescent, male, and slightly less educated (Figure 6).

Figure 6. Sentiment analysis & profiling.

In prior work, Pennebaker (2011) demonstrated how personal writing style contains subtle cues for age, gender, state-of-mind, and personality. For example, statistically, women tend to use more pronouns (*I, my, we*) to talk about relationships, while men tend to use more determiners and quantifiers to talk about objects and concepts (*the, one, more, …*), and adolescents use more informal language and profanity, while introverted persons use more negative adjectives.

The results indicate that 35% of the users are women, likely because most of the discourse is about interpersonal relationships and physical functions, topics that are statistically more prevalent with women (Newman et al. 2008, 219). It could be that the API is correct, and that there really are 35% women participating on Incels.me (e.g., fake profiles,) but given the level of aggression we consider this chance slight.

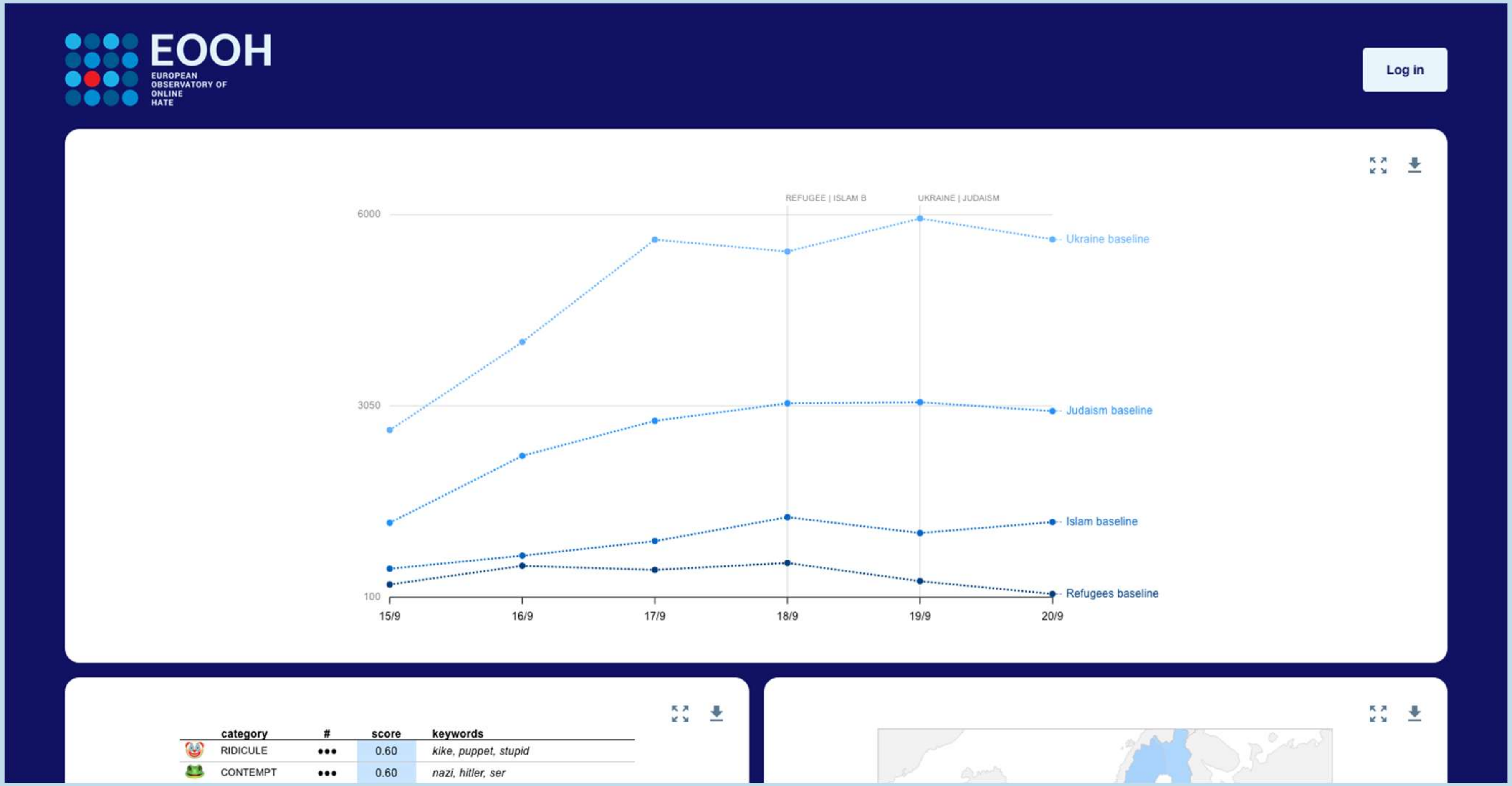
Younger age correlates with lower education, suggesting that the users are not necessarily less educated than other people, although there are multiple references to *low IQ* (~50x), which may indicate low self-esteem (or contempt).

[13] "Textgain Real-time Text Analytics", https://www.textgain.com

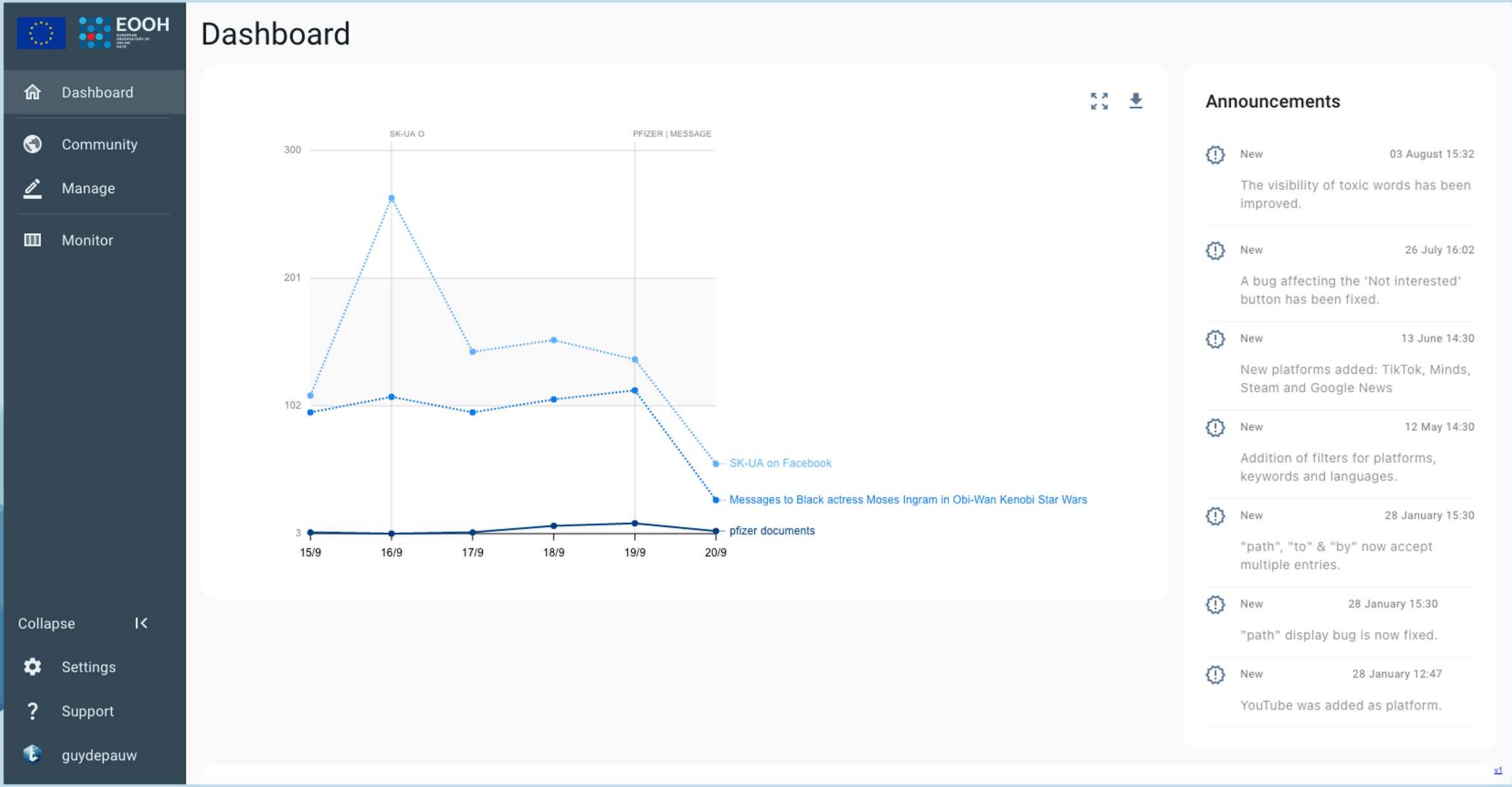
QAnon: researchgate.net/publication/The-QAnon-superconspiracy
Islamic State: arxiv.org/pdf/1803.04596.pdf
German far-right: organisms.be/downloads/jaki2018.pdf
Incels: organisms.be/downloads/incels.pdf

DASHBOARD

European Observatory of Online Hate



	category	#	score	keywords
🤡	RIDICULE	●●●	0.60	*kike, puppet, stupid*
🐸	CONTEMPT	●●●	0.60	*nazi, hitler, ser*



Dashboard

Announcements

New 03 August 15:32

The visibility of toxic words has been improved.

New 26 July 16:02

A bug affecting the 'Not interested' button has been fixed.

New 13 June 14:30

New platforms added: TikTok, Minds, Steam and Google News

New 12 May 14:30

Addition of filters for platforms, keywords and languages.

New 28 January 15:30

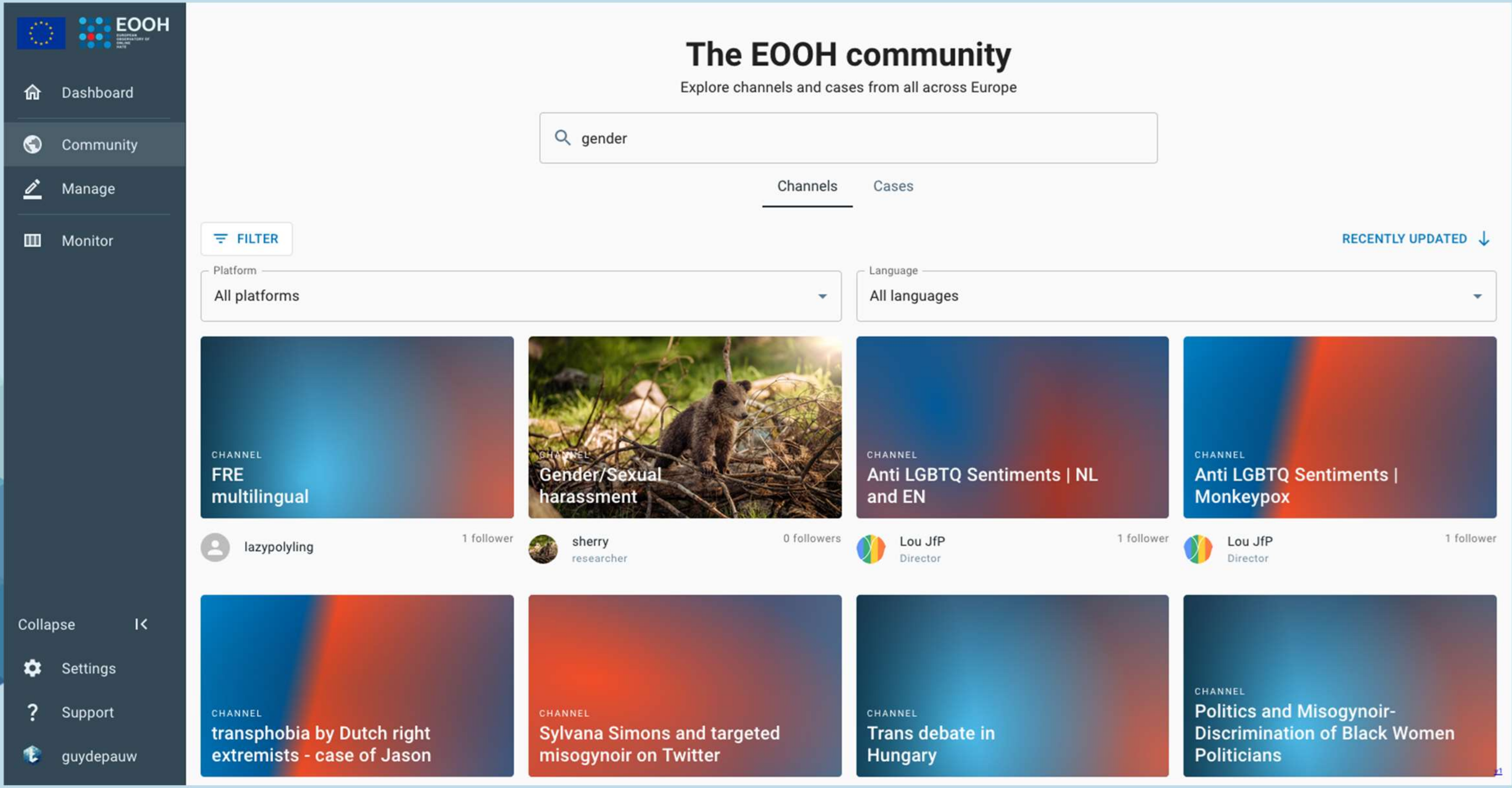
"path", "to" & "by" now accept multiple entries.

New 28 January 15:30

"path" display bug is now fixed.

New 28 January 12:47

YouTube was added as platform.



The EOOH community

Explore channels and cases from all across Europe

gender

Platform: All platforms

Language: All languages

CHANNEL
FRE multilingual
lazypolyling — 1 follower

CHANNEL
Gender/Sexual harassment
sherry, researcher — 0 followers

CHANNEL
Anti LGBTQ Sentiments | NL and EN
Lou JfP, Director — 1 follower

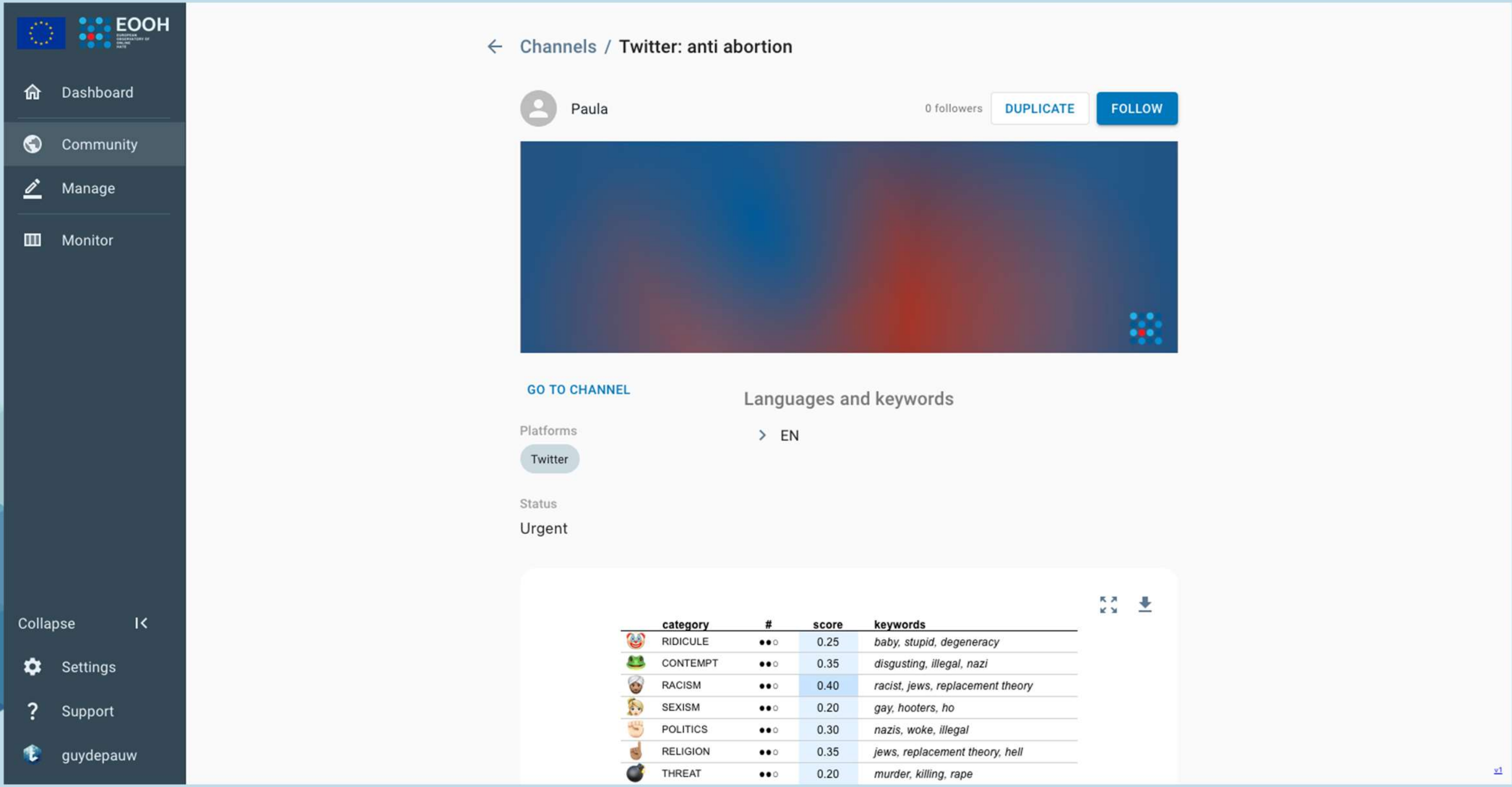
CHANNEL
Anti LGBTQ Sentiments | Monkeypox
Lou JfP, Director — 1 follower

CHANNEL
transphobia by Dutch right extremists - case of Jason

CHANNEL
Sylvana Simons and targeted misogynoir on Twitter

CHANNEL
Trans debate in Hungary

CHANNEL
Politics and Misogynoir-Discrimination of Black Women Politicians



Channels / Twitter: anti abortion

Paula

0 followers

Platforms

Twitter

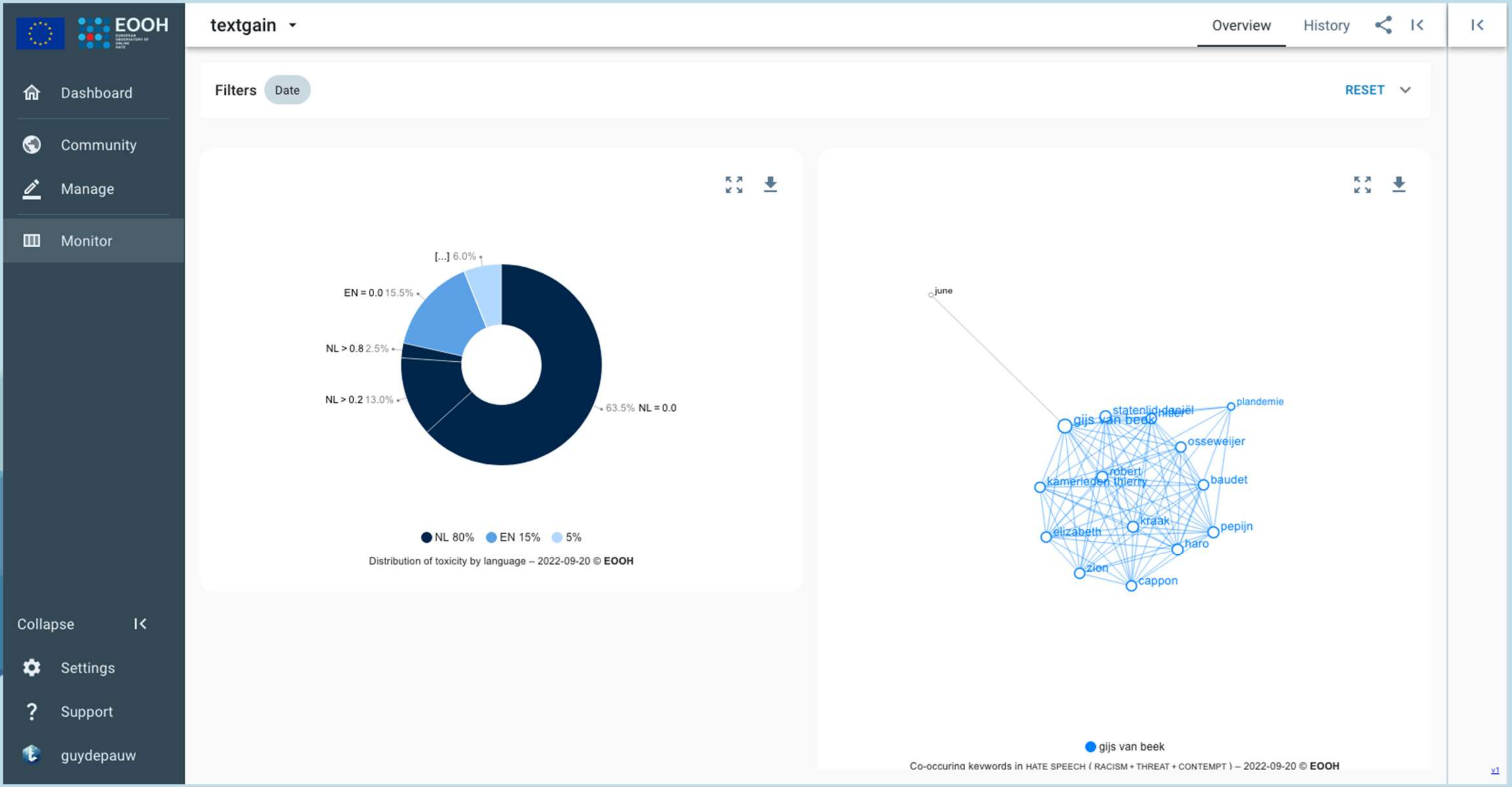
Status

Urgent

Languages and keywords

EN

	category	#	score	keywords
🤡	RIDICULE	●●○	0.25	*baby, stupid, degeneracy*
🐸	CONTEMPT	●●○	0.35	*disgusting, illegal, nazi*
👳🏽	RACISM	●●○	0.40	*racist, jews, replacement theory*
👱‍♀️	SEXISM	●●○	0.20	*gay, hooters, ho*
✊🏻	POLITICS	●●○	0.30	*nazis, woke, illegal*
☝🏽	RELIGION	●●○	0.35	*jews, replacement theory, hell*
💣	THREAT	●●○	0.20	*murder, killing, rape*



Distribution of toxicity by language – 2022-09-20 © **EOOH**

Co-occuring keywords in HATE SPEECH (RACISM + THREAT + CONTEMPT) – 2022-09-20 © **EOOH**

Coming up…

- User manual
- Speed boosts through better hardware
- Secure email pipelines for automated email alerts and weekly digests
- Trending words and phrases

Judaism

- *hassidic zionist*
- *join telegram*
- *conquer nwo*
- *exposed reptilian*
- *gassed millions*
- *politics demonic*
- *starved camps*
- *systematic effort*
- *camps typhus*
- *camps de-loused*

Islam

- *shirk*
- *insult lgbtq+*
- *global agenda*
- *global fascism*
- *jewish massacres*
- *abused girls*
- *men jailed*

Refugees

- *grateful ukr*
- *good kills*
- *independence day*
- *battalion delivering*
- *ukie escape*
- *national army*
- *chechens trophies*
- *malkachan love*
- *globohomo art*

Ukraine

- *#crimea*
- *taiwan*
- *western globalist*
- *sovereign development*
- *zaporizhzhia nuclear*
- *#visabanforrussians*
- *nuclear plant*
- *visa ban*

Coming up…

- Theme taxonomy
  - 118 experts
  - 727 channels (532 active)
  - Develop a living theme taxonomy based on the interests of the experts
    - Hands-on pan-European trend analysis

- Local upload of data for analysis

Dataset description

- The dataset has over 550K messages (*n*=558,918).
- These were posted between July 01, 2022 and August 31, 2022.
- Most data was collected from Twitter (80%), Facebook and 4plebs.
- Most messages are written in English (EN, 65%), Dutch (NL, 10%) and French (FR, 5%).
- The **average toxicity score** is **0.35**, which is quite high.
- About 50K messages (10%) are very toxic, with a score > 0.8.
- Some of these are written in English (EN, 55%).
- For example: *“>>391799441 >Chinks vs kikes Chinks are nothing but kike slaves. Globohomo wouldn’t be able to exist without chink slave labor. If you want to defeat the Jews, then china must be destroyed.”* (source: 4plebs).
- Messages that are **very toxic** (> 0.8) are more frequent on Twitter (35%).
- Toxicity often involves RACISM (10%), POLITICS, RELIGION and/or CONTEMPT.
- About 5% of messages involve THREATENING.
- About 2% of messages involve DISINFORMATION.
- The most frequent combination is RACISM + RELIGION.
- Some common toxic keywords include: *holohoax*, *holokauszt*, *nagyon*, *tant*, *scum*, *jews control*, *jews hate*, *latvijā*, *ježiš*, and *shitskin*.

JOIN US AND BECOME A EOOH PARTNER!

Q&A

textgain
Textgain
Olivier Cauberghs
Olivier@Textgain.com