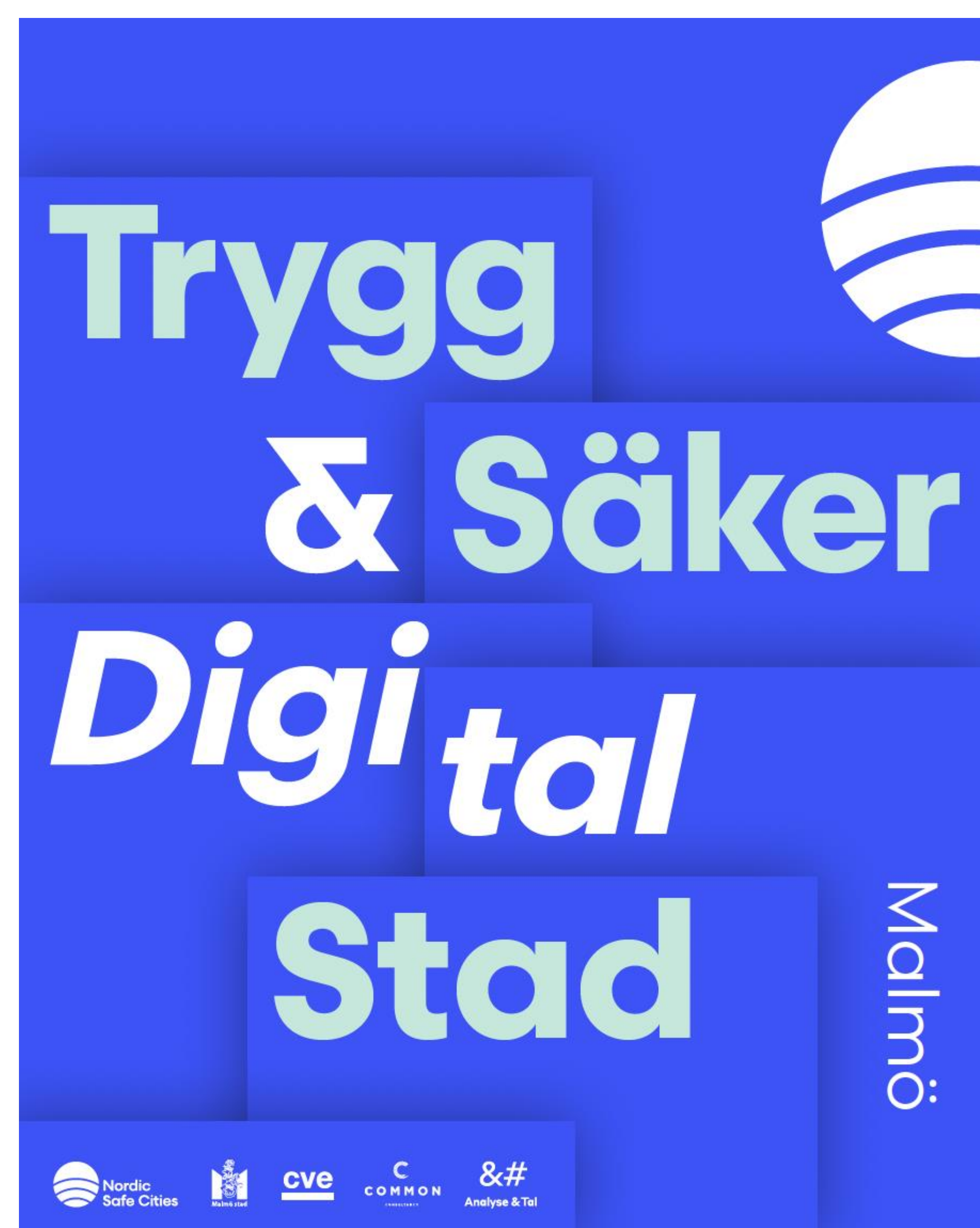




Tryg Digital By

Tryg Digital By



Niels Ørbæk Chemnitz, *Analyse & Tal*

Sebastian Jørgensen, *Nordic Safe Cities*



Agenda

- 1 Projektet: Lidt baggrund**
- 2 Sproglige angreb på dansk: Fra idé til model**
- 3 Fra dansk til norsk: En tvær-skandinavisk model**
- 4 Resultater: Hvad kan vi bruge det til?**



Projektet
Lidt baggrund

The background of the slide is a photograph of a city street at sunrise or sunset. The sun is low on the horizon, creating a warm, orange glow. The street is lined with trees and parked cars. Overlaid on the image are large, stylized symbols: a black ampersand (&) in the top left, a black hash (#) in the top right, a large orange ampersand (&) in the middle left, a large orange hash (#) in the middle right, and a large white ampersand (&) in the bottom left and a large white hash (#) in the bottom right.

Analyse & Tal f.m.b.a.

Analyse & Tal er et digitalt analysebureau med kontorer i **København, Århus** og **Oslo**. Vi er specialiserede i at analysere dét, der kan være svært at analysere.

Analyse & Tal har eksisteret siden 2014 og tæller i dag omkring **25 faste medarbejdere**. Vores drøm er at skabe et mere **demokratisk og retfærdigt samfund**. Derfor har vi valgt at organisere os som et **medarbejderejet kooperativ**.

Vi er stolte af at investere vores overskud i **udviklingen af nye metoder**, projekter og i demokratiseringen af vores samfund som helhed.



"To ud af tre danskere afholder sig i dag fra at debattere på sociale medier, fordi tonen er for hård"

- TV2, Megafon, 2020

Men hvor hård er tonen egentlig? Er den blevet hårdere? Hvem står for skud? Og hvilke emner antænder debatten?



**Sproglige angreb
på dansk
Fra idé til model**

Datagrundlag:

Hvordan afgrænser man den offentlige debat på sociale medier?

Mediesider

54.492.711

Kommentarer under

994.466

opslag fra

477

Pressenævnstilmeldte mediers
Facebooksider*

Politikersider

8.583.465

kommentarer under

102.646

Opslag fra

199

Folkevalgte politikere og partiers
Facebooksider

Sådan har vi defineret fænomenerne:

Angreb

Stigmatiserende, nedsættende, krænkende, stereotypiserende, ekskluderende, chikanerende eller truende ytringer

Herunder:

- **Hadtale**

Angreb mod en gruppe eller individ baseret på dennes beskyttede karakteristika:

- **Anstødelige angreb**

Angreb mod en gruppe eller et individ, der ikke er baseret på personens eller gruppens beskyttede karakteristika.

OBS: Vi er kun interesserede i sproglige angreb, ikke alt anstødeligt sprog som fx bandeord.¹

Beskyttede karakteristika

Race/etnicitet, hudfarve, nationalitet og oprindelse, religion og tro, seksuel orientering, køn og kønsidentitet, socialklasse og social status, politisk orientering, alder eller handicap og seriøse sygdomme (både fysiske og psykiske).

[1] Sigurbergsson, G. I., & Derczynski, L. (2019). Offensive Language and Hate Speech Detection for Danish (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1908.04531>

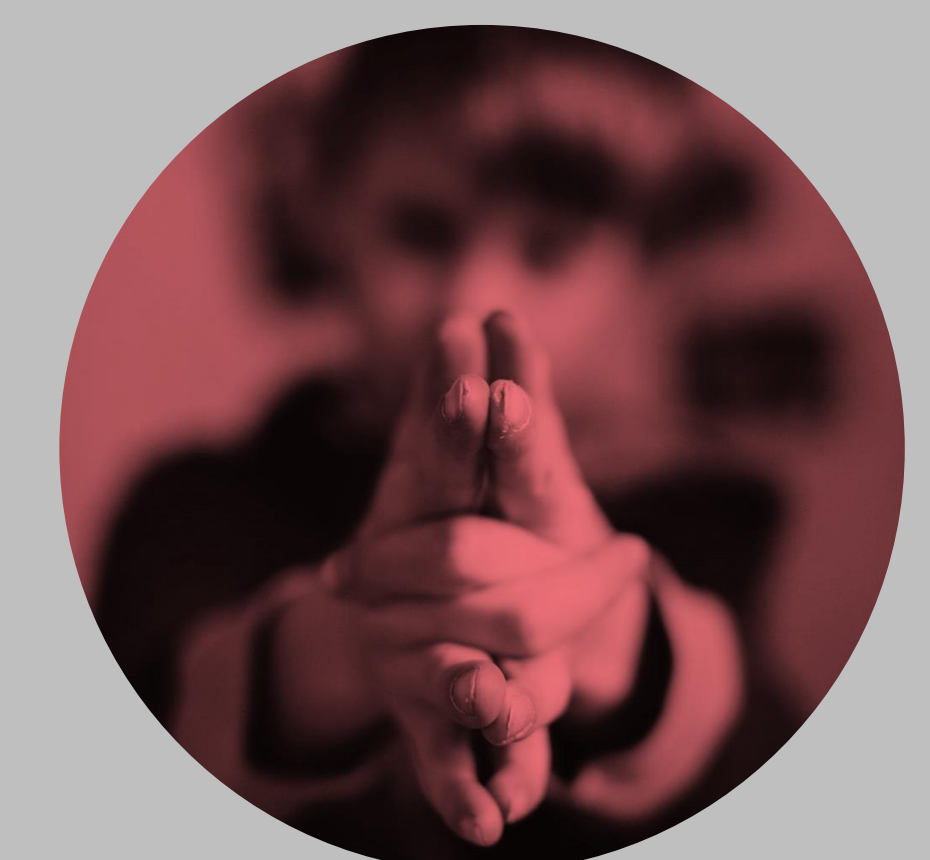
Angreb vs. ikke angreb



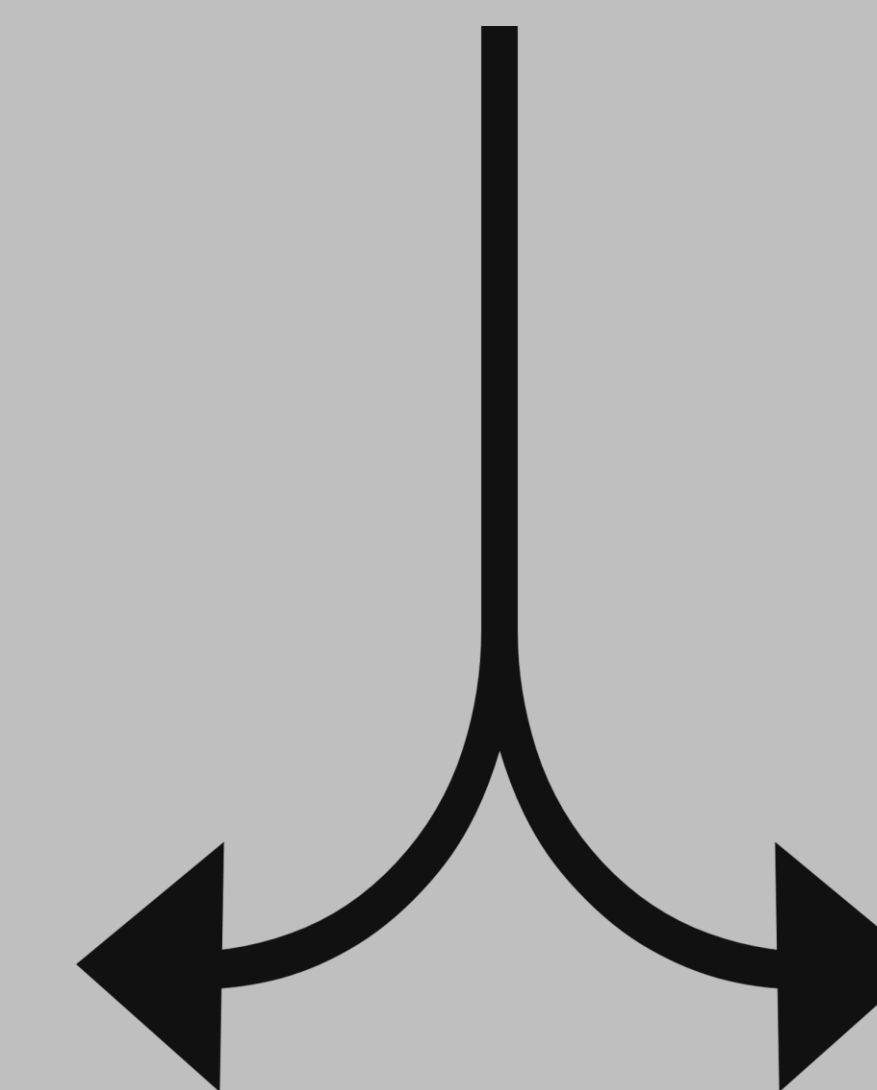
ANGREB



ANSTØDELIGHED



HADTALE



Annotering af data

“How training datasets are annotated is one of the most important aspects of their creation” - Vidgen and Dercynski, 2020¹

Udvikling af udførlig kodemanual og løbende opkvalificering af annotører.

Vi ansatte 4 annotører med i alt ~600 timer til rådighed og opnåede ~80,000 annoterede kommentarer, hvoraf størstedelen var til [A&ttack](#), [Ha&te](#) og [Rec&nition](#)

[1] Vidgen, B., & Derczynski, L. (2020). Directions in Abusive Language Training Data: Garbage In, Garbage Out. arXiv. <https://doi.org/10.48550/ARXIV.2004.01670>

Indeholder kommentaren angreb, anerkendelse, begge dele eller ingen af delene?

Ned i havnen med noget tungt om benene med den møgunge, du er her ikke for sjov, eller for dine børn, du er for at skabe profit til arbejdsgiverne. Ellers kan de jo ikke nyde deres afgiftslettelser og skattelettelser som DF har kæmpet så hårdt for. Egoistiske tøs.

☐ Angreb (anstødelig eller hadefuld retorik) 1

☐ Anerkendelse 2

☐ Både angreb og anerkendelse 3

☐ Hverken angreb eller anerkendelse 4

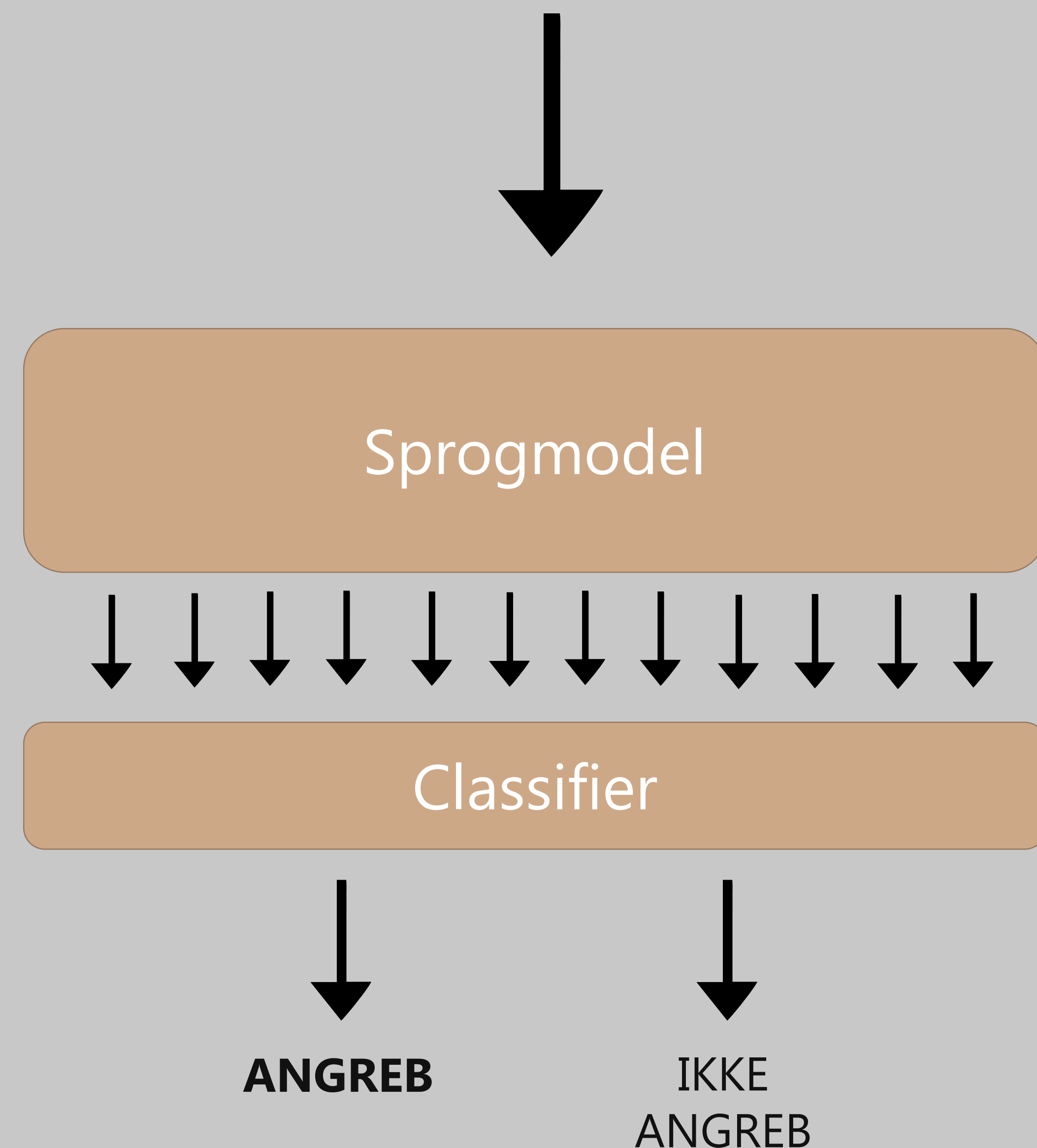
☐ Svært at sige 5

☐ Konteksten var af afgørende betydning for min vurdering 6

☒
☐
☐

Sprogmodeller og classifilers

“Hun ter sig som en Sindssyg kælling!!! Hun tænker ikke på de mange tabte arbejdspladser”



Grundlaget for vores **classifler**, der vil inddele tekstbidder i "angreb" eller "ikke angreb" er en **sprogmodel** der er trænet på en **stor mængde tekst**.

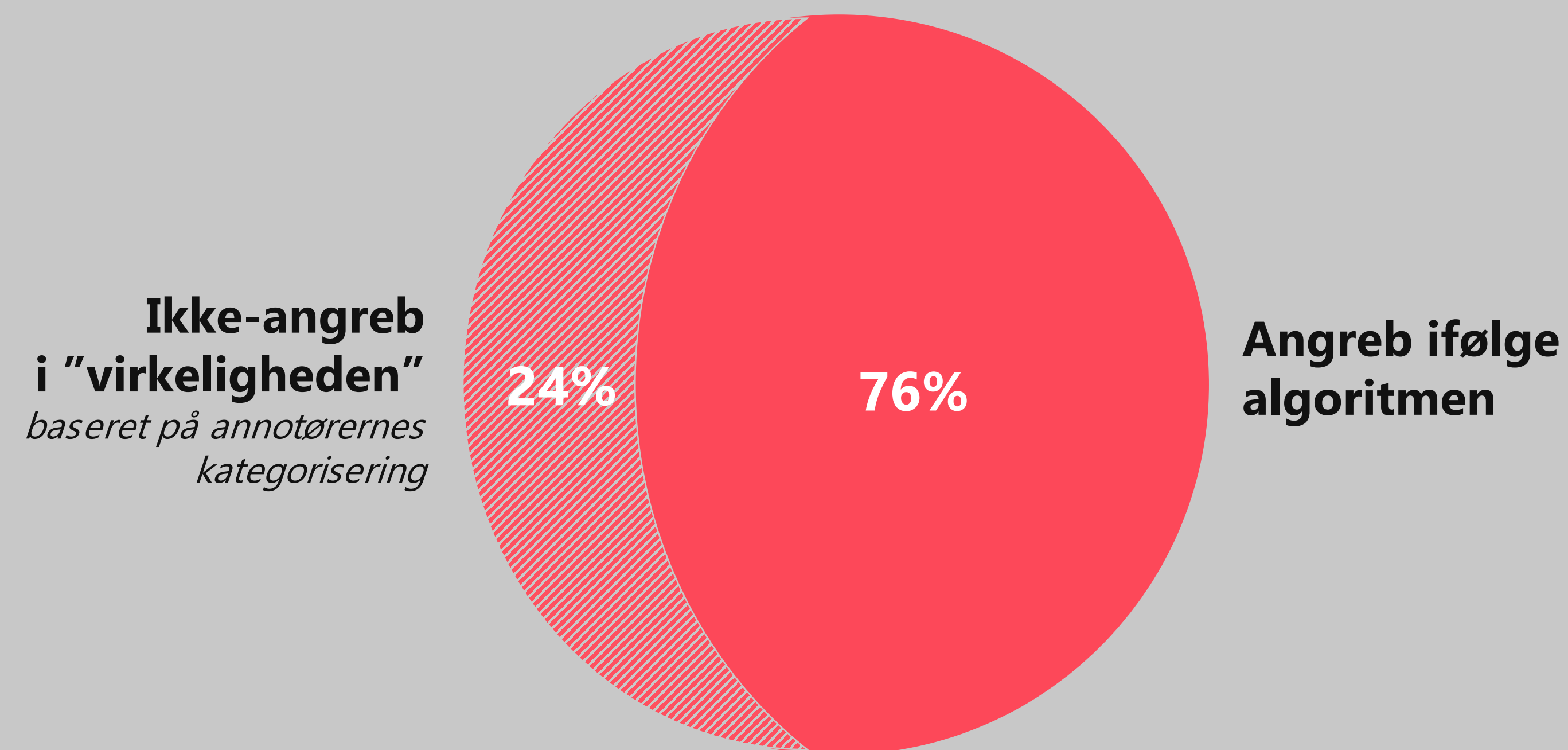
Sprogmodellen **encoder** en tekstbid i et mange-dimensionelt vektorrum og spytter en enkelt vektor ud der består af en lang række tal.

Vores classifler **lærer** fra disse vektorer og prøver at finde **et genkendeligt mønster** der adskiller angreb og ikke angreb.

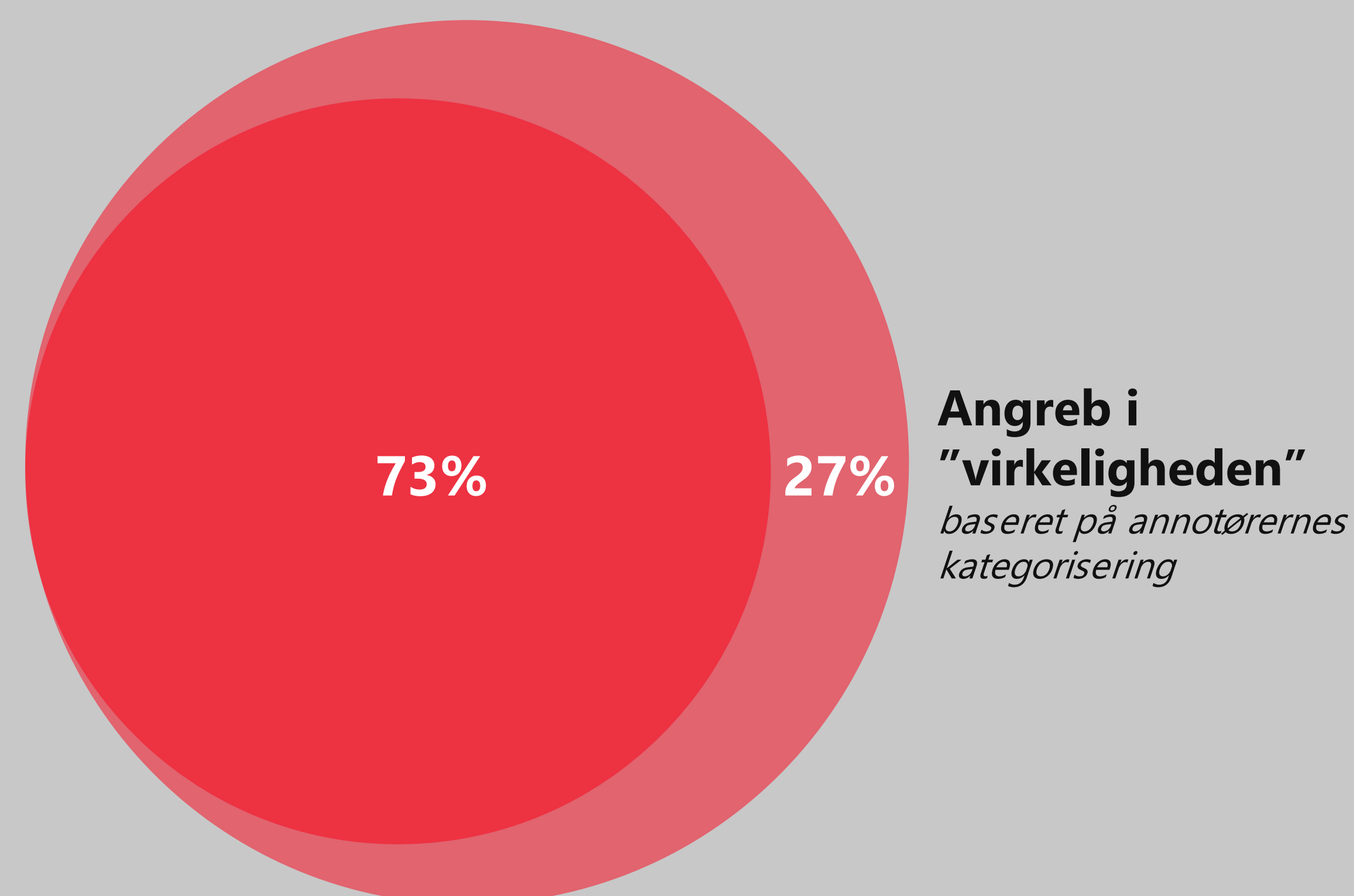
Vi brugte **Ælæctra**, en dansk model på Electra arkitekturen, som sprogmodel. Denne model er trænet på tekstsamlingen **Danish Gigaword**

Algoritmens formodede usikkerhed

Estimat af andelen af ikke-angreb i
algoritmens output af angreb



Estimat af manglende angreb i
algoritmens output af angreb



Macro avaraged F1-score:

Angreb/ikke angreb (At&tack) : **0.8341**

Had vs. anstødelighed (Ha&te): **0.8171**

Træning og evaluering af A&ttack

Test-datasæt:

Den endelige algoritme evalueres på 10% af datasættet der ikke er blevet brugt til udviklingen af modellen

Makro F1-score: En metrik der afvejer *precision* og *recall*.

På linje med de bedste:

Algoritmen opnåede en makro F1 på **0.83**, hvilket var det bedste i Danmark og på linje med de bedste på høj-ressource sprog som engelsk.



Fra dansk til norsk
En tvær-
skandinavisk
model

Sproglige angreb på norsk

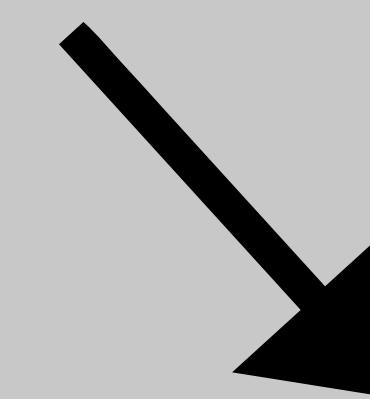
Inspireret af succesen med A&ttack, satte vi os for at lave en tilsvarende **norsk algoritme**, men løb hurtigt ind i to nye problemer.

1. Flere skriftsprog:

Norsk har **to officielle skriftsprog**, bokmål og nynorsk, og et hav af dialekter der også kommer til udtryk skriftligt. Dette gør opgaven betydeligt sværere.

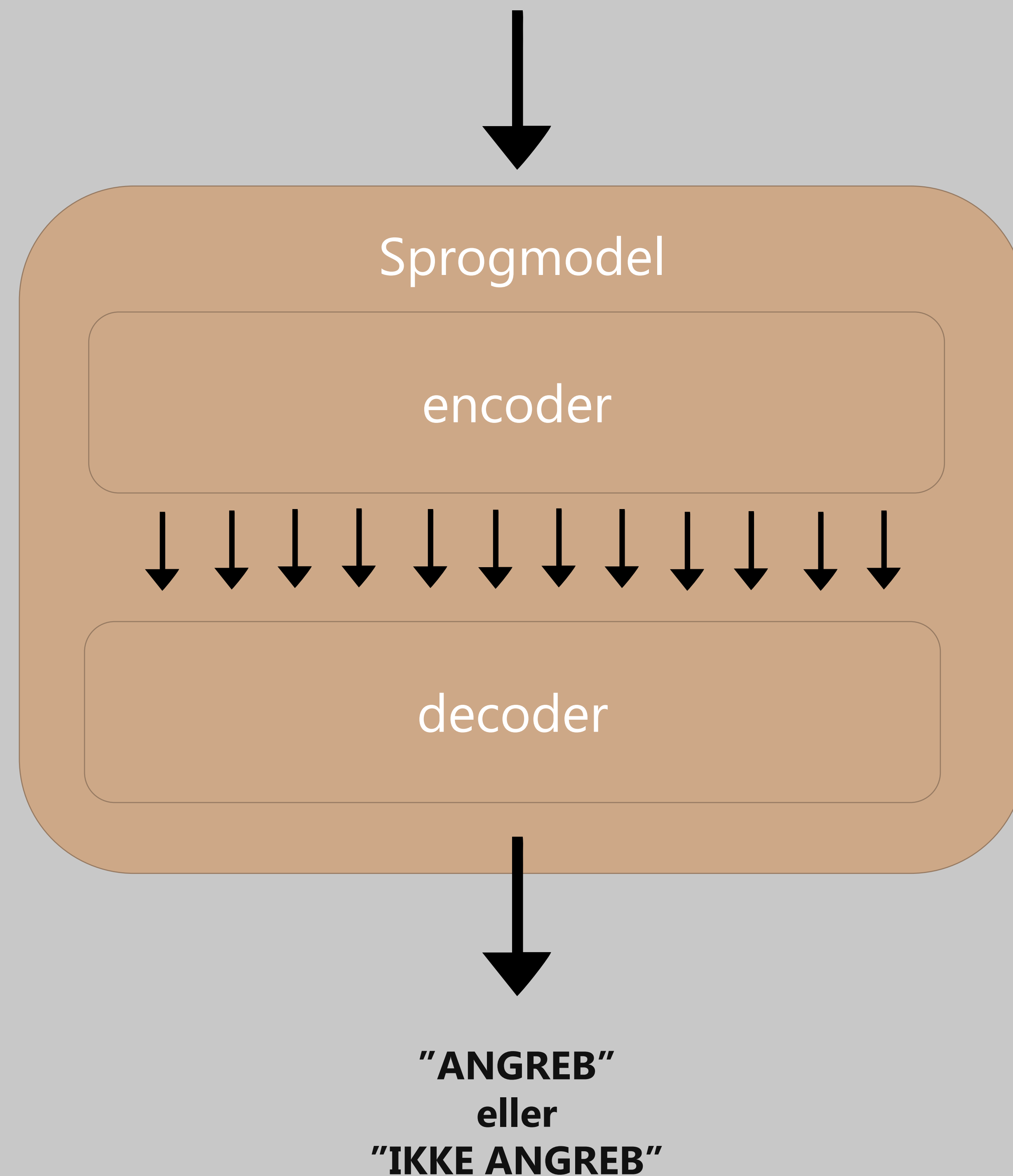
2. Sprogmodel og korpus:

Der fandtes ikke en tilsvarende sprogmodel, og vi måtte derfor eksperimentere med en række andre arkitekturer. Tekstsamlingen **Norwegian Colossal Corpus** er større end Danish Gigaword, men indeholder ikke noget social medie data.



Sequence-to-Sequence Models

"Hun ter sig som en Sindssyg kælling!!! Hun tænker ikke på de mange tabte arbejdspladser"



T5-sprogmodeller

Vi afprøvede mange forskellige typer af sprogmodeller, og de der fungerede bedst var af familien **T5**: Text-to-Text Transfer Transformer.

Disse modeller tager en tekst-bid som input og giver en tekst-bid som output, og egner sig specielt til opgaver som oversættelse, summering, spørgsmålsbesvarelse, osv. Men de kan også bruges til fx. klassifikation

T5 modellerne findes i flere størrelser, hvor de større modeller øger kompleksiteten af modellen og kravene til det hardware der skal træne/bruge den.

Sprogvarianter

T5-modeller findes i mange variationer. Den tekstsamling de er blevet pretrænet på danner grundlag for deres behandling af sprog. Her præsenterer vi resultater fra 3 sprogvarianter.

- En norsk model
- En dansk model
- En skandinavisk model

Alle modellerne er størrelse *large*, hvilket vil sige at de har 770 millioner parametre.



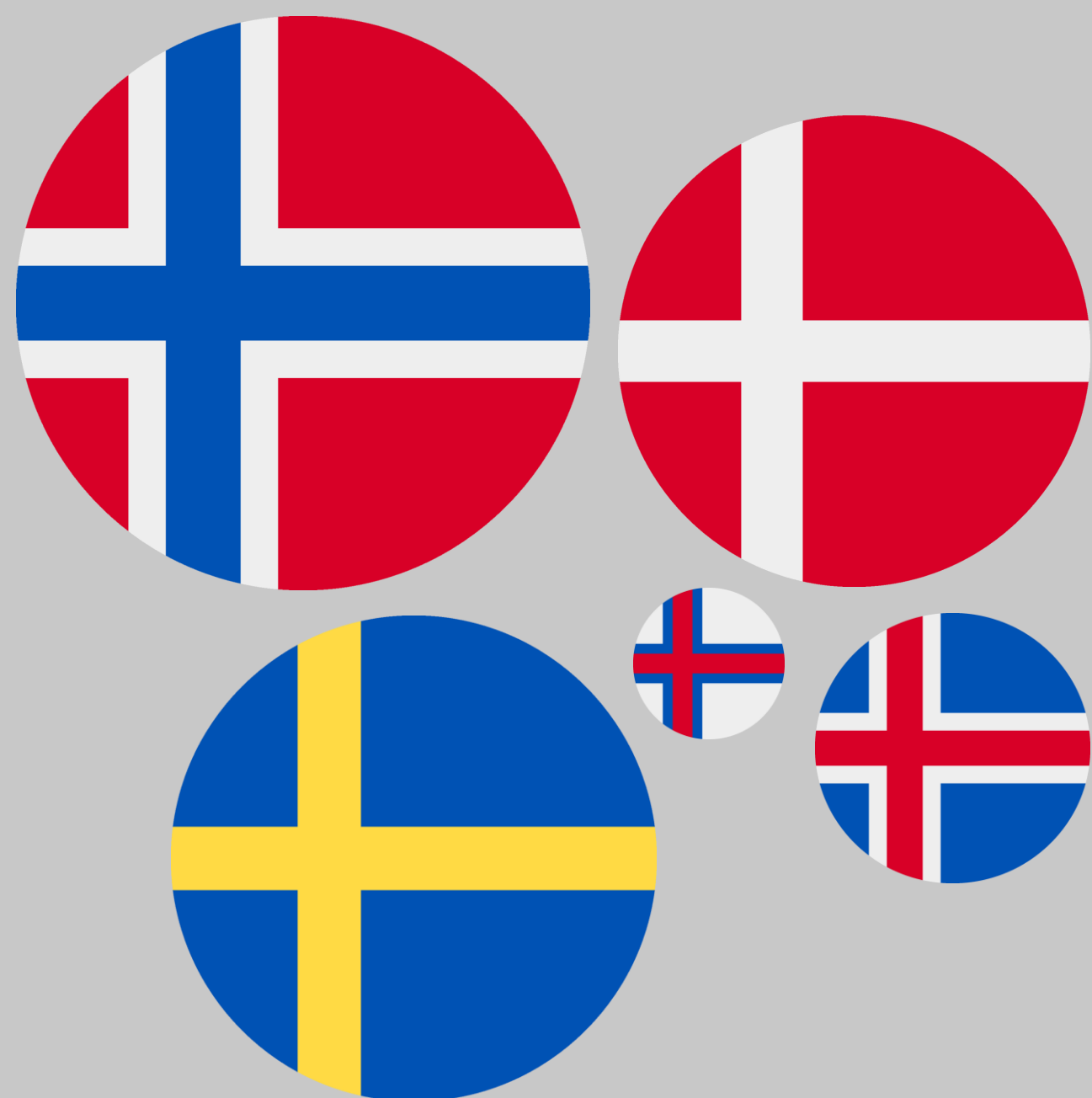
north/t5_large_NCC

En norsk T5 large trænet på Norwegian Colossal Corpus. Udviklet af Per E Kummervold



strombergnlp/dant5-large

En dansk T5 large trænet på Danish GigaWord. Udviklet af StrømbergNLP ved ITU



north/t5_large_scand3M

En eksperimentiel skandinavisk T5 large trænet på en tekstsamling bestående af norsk, dansk, svensk, islandsk og en smule færøsk. Udviklet af Per E Kummervold

Evaluering

- Den skandinaviske sprogmodel klarede sig bedre på begge delopgaver (dansk og norsk)
- Forbedring på 2-3 procentpoint i forhold til enkeltsprogmodeller
- At skulle håndtere to sprog burde være sværere, ikke?
- Men fordelene ved et større datasæt virker til at trumfe dette
- Næste skridt er at tilføje svensk og islandsk data.

<i>Sprogmodel</i>	<i>Norsk</i>	<i>Dansk</i>	<i>Data</i>
Norsk T5 large*	0.75		Norsk
Skandinavisk T5 large**	0.78	0.87	Norsk+Dansk
Dansk T5 large***		0.85	Dansk
A&ttack		0.83	Dansk

Alle metrikker er F1-makro

* https://huggingface.co/north/t5_large_NCC
** https://huggingface.co/north/t5_large_scand3M (begrænset adgang)
*** <https://huggingface.co/strombergnlp/dant5-large>

4

4

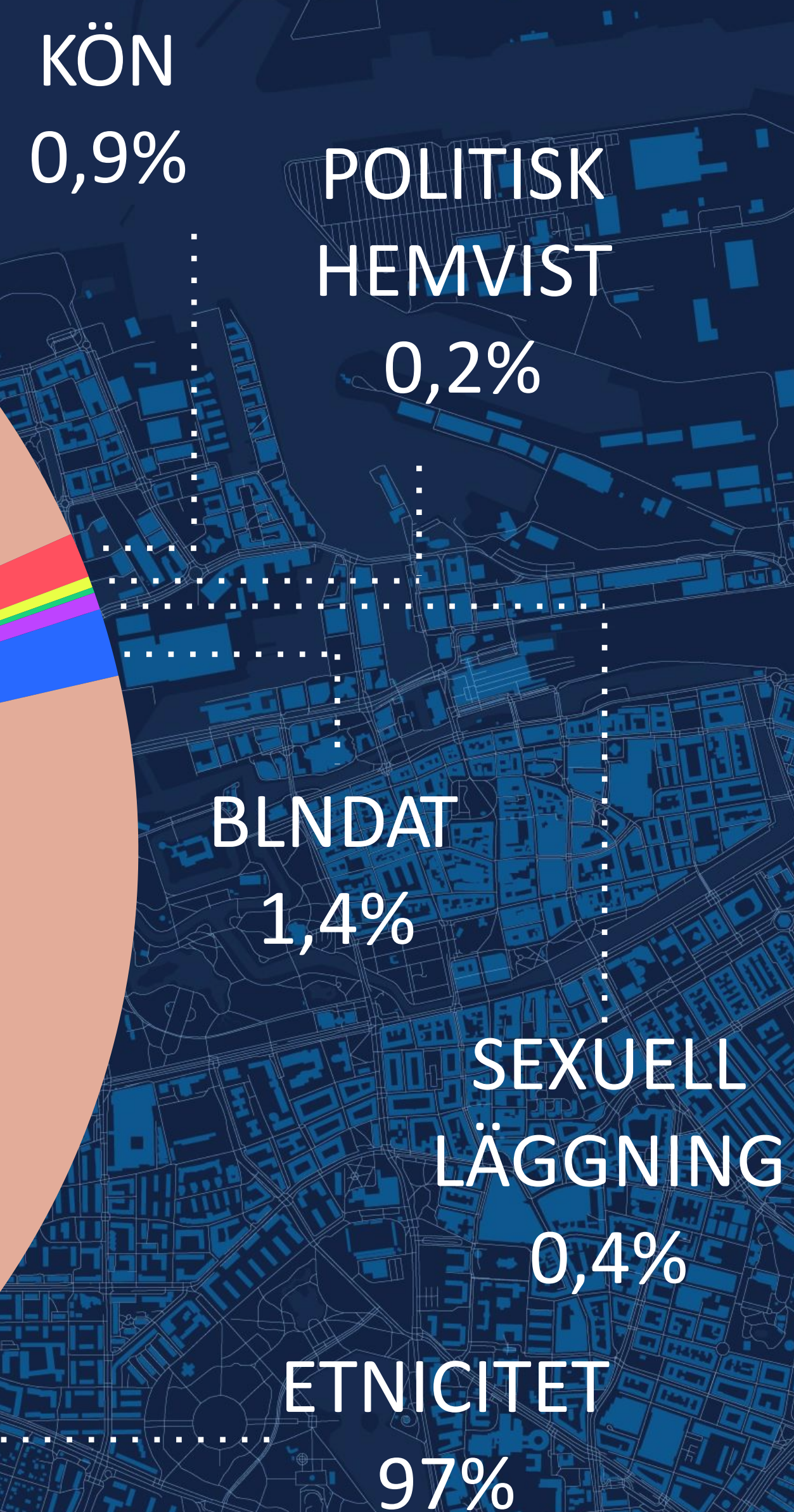
4

4

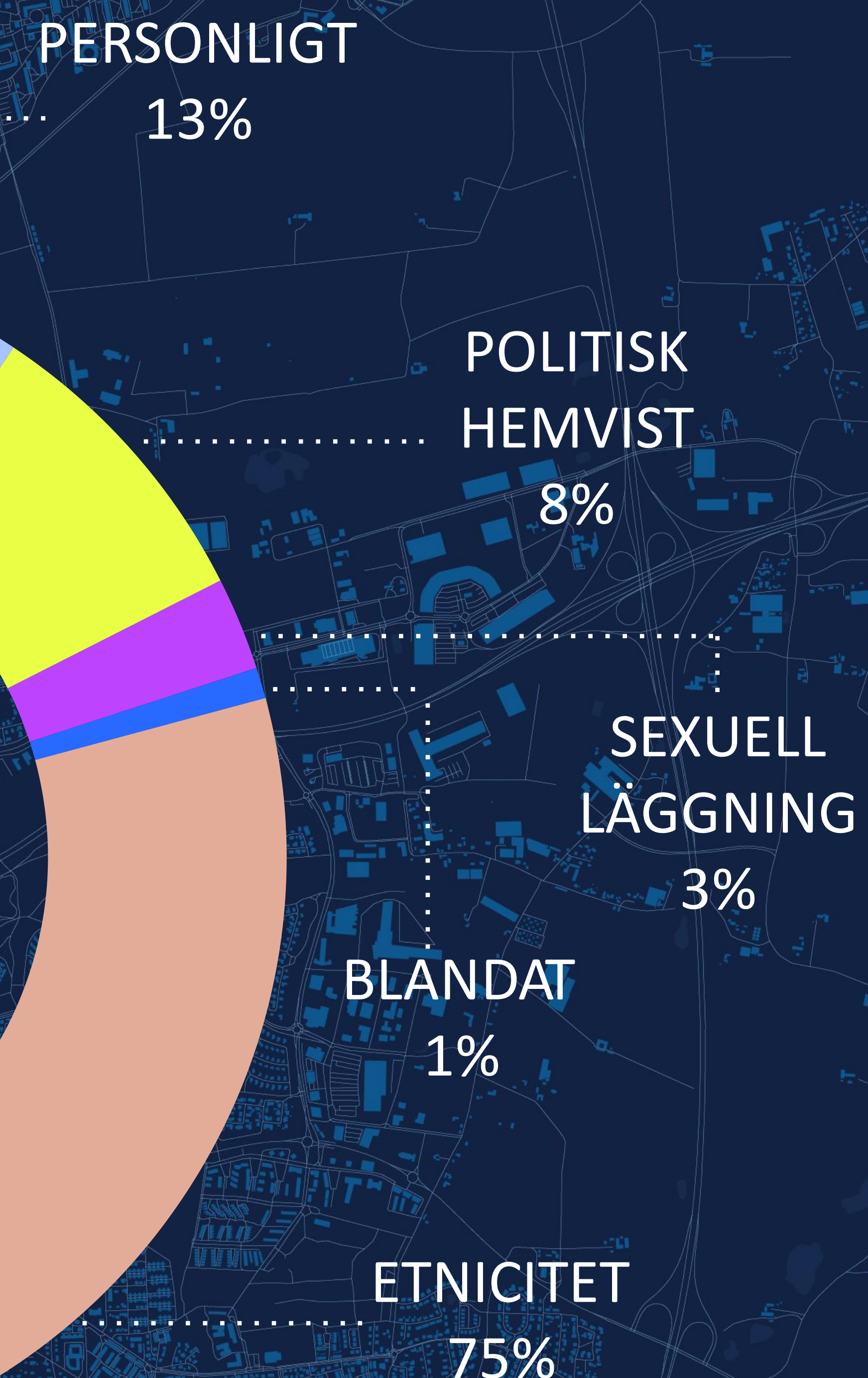
Resultater
Hvad kan vi
bruge det til?

Næsten alt hadet er rettet mod etniske minoriteter.
Vi deles ikke lige om hadet.

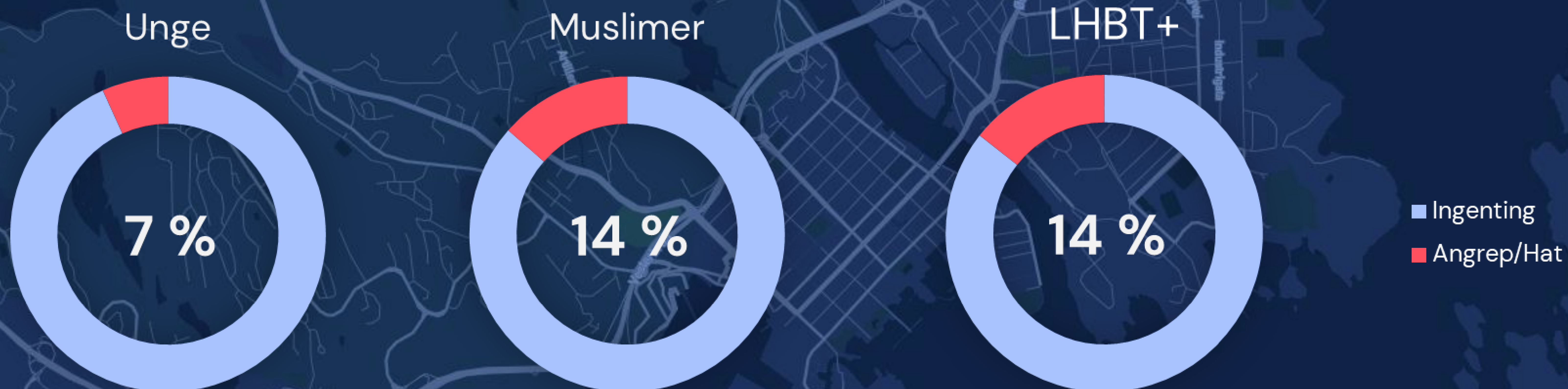
FLASHBACK



FACEBOOK

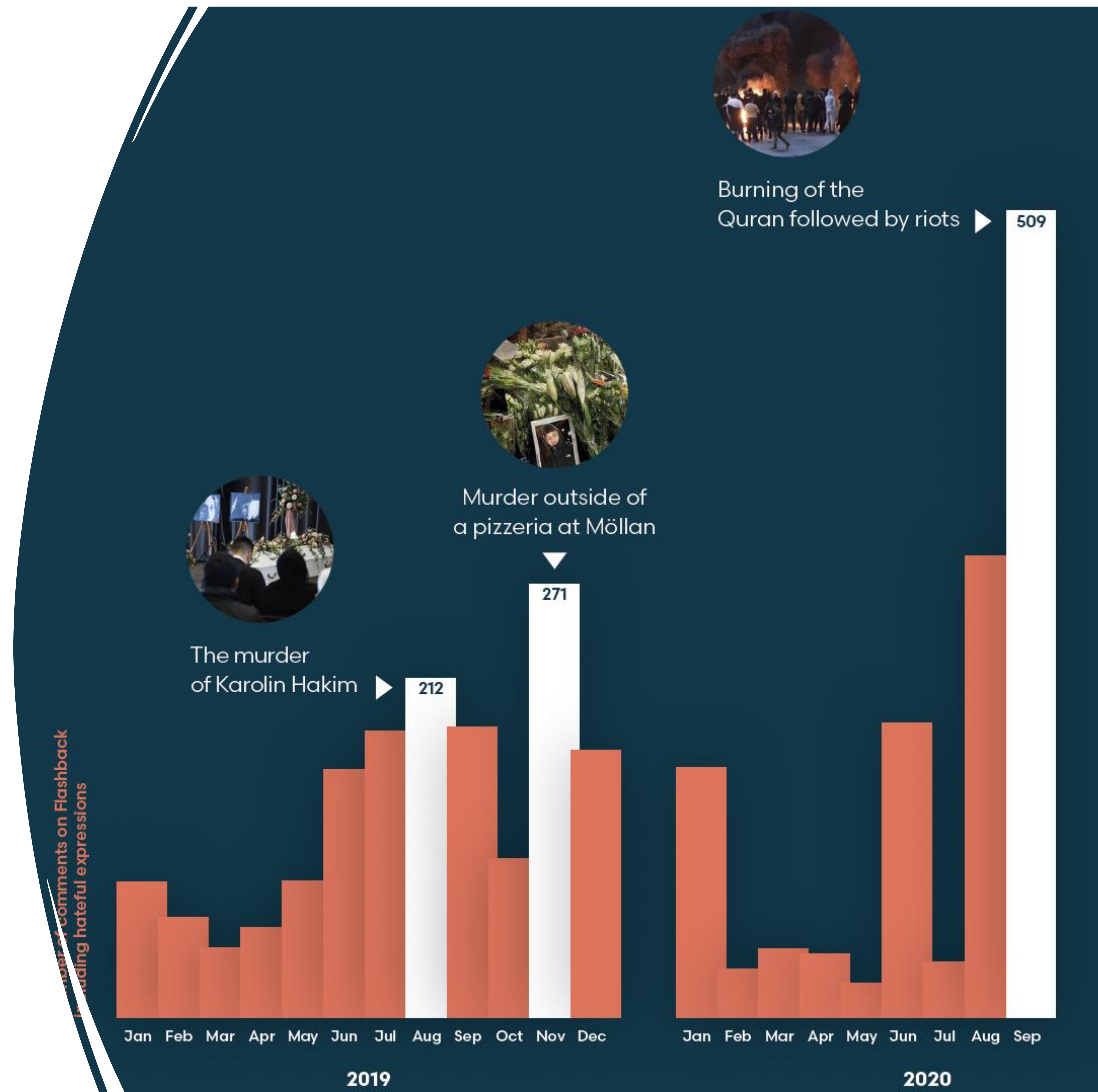


Men når samtalen eksplicit handler om etniske minoriteter, LGBTQ+ personer eller unge, er andelen angreb og hat høj



Fysiske rum & digitale rum

- Hadet eksisterer i byen og er ikke begrænset til enten digitale eller fysiske rum.





Tryg Digital By

3 Digitale veje frem i Malmö



1. The City of Malmö and the Police.

A digital safety team

With the ambition to prevent, combat and address polarisation in the digital debate, in the long-term as well as during acute crises.



2. Civil society: A digital intervention network

With the ambition to strengthen the digital voice of the city's civil society organisations so that they can contribute to online safety, to a more nuanced debate and to a safe local online democracy.



3. Elected officials

With the ambition to strengthen the ability to take action for a safe and secure local democracy and a non-polarising and inclusive debate climate.

Tak for
opmærksomheden

